ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2018)
67, *Part* 5, *pp.* 1275–1304

Unravelling the predictive power of telematics data in car insurance pricing

Roel Verbelen,

KU Leuven, Belgium

Katrien Antonio

KU Leuven, Belgium, and University of Amsterdam, The Netherlands

and Gerda Claeskens

KU Leuven, Belgium

[Received January 2017. Revised March 2018]

Summary. A data set from a Belgian telematics product aimed at young drivers is used to identify how car insurance premiums can be designed based on the telematics data collected by a black box installed in the vehicle. In traditional pricing models for car insurance, the premium depends on self-reported rating variables (e.g. age and postal code) which capture characteristics of the policy(holder) and the insured vehicle and are often only indirectly related to the accident risk. Using telematics technology enables tailor-made car insurance pricing based on the driving behaviour of the policyholder. We develop a statistical modelling approach using generalized additive models and compositional predictors to quantify and interpret the effect of telematics variables on the expected claim frequency. We find that such variables increase the predictive power and render the use of gender as a rating variable redundant.

Keywords: Compositional predictors; Generalized additive models; Pay as you drive insurance; Risk classification; Structural 0s; Usage-based insurance

1. Introduction

For a unique Belgian portfolio of young drivers in the period between 2010 and 2014, telematics data on how many kilometres are driven, during which time slots and on which type of roads were collected using black box devices installed in the insured drivers' cars. Our aim is to incorporate this information in statistical rating models, where we focus on predicting the number of claims, to set premium levels adequately on the basis of individual policyholder's driving habits.

Determining a fair and correct price for an insurance product (also called *rate making*, *pricing* or *tarification*) is crucial for both insured drivers and insurance companies. Car insurance is traditionally priced on the basis of self-reported information from the insured person, most importantly age, licence age, postal code, engine power, use of the vehicle and claims history. However, these observable risk factors are only proxy variables, not reflecting current driving habits and driving style. Telematics technology—the integrated use of telecommunication and informatics—may fundamentally change the car insurance industry. The use of this technology

Address for correspondence: Roel Verbelen, ORSTAT, KU Leuven, Naamrestraat 69, Bus 3555, 3000 Leuven, Belgium.
E-mail: roel.verbelen@kuleuven.be

in insured vehicles enables the transmission and receipt of information that allows an insurance company to quantify the accident risk of drivers better and to adjust the premiums accordingly through usage-based insurance (UBI). By monitoring their customers' motoring habits, underwriters can increasingly distinguish between drivers who are safe on the road from those who merely seem safe on paper (Economist, 2013). Young drivers and drivers in other high risk groups, who typically face hefty insurance premiums, can be judged on the basis of how they really drive. Regulation also plays a role as the use of indirect indicators of risk is being questioned by the European Court of Justice. In 2012, a European Union (EU) ruling came into force, banning price differentiation based on gender (`http://europa.eu/rapid/press-release_IP-11-1581_en.htm`). Through telematics, women may be able to confirm that they really are safer drivers.

The use of telematics risk factors potentially enables an improved method for determining the cost of insurance. Because of more refined customer segmentation and greater monitoring of driving behaviour, UBI addresses the problems of adverse selection and moral hazard that arise from the information asymmetry between the insurer and the policyholders (Filipova-Neumann and Welzel, 2010). Closer aligning insurance policies to the actual risks increases actuarial fairness and reduces cross-subsidization compared with grouping drivers into too general actuarial classes (Desyllas and Sako, 2013). Telematics insurance gives a high incentive to change the current driving pattern and stimulates more responsible driving (Parry, 2005; Litman, 2015; Tselentis *et al.*, 2016). Users' feedback on driving behaviour and game design elements in UBI can further enhance the customer experience by making it more interactive, gratifying and even exciting (Toledo *et al.*, 2008). Less and safer driving is encouraged, leading to improved road safety and reduced vehicle travel with less congestion, pollution, fuel consumption, road cost and crashes (Greenberg, 2009).

Usage-based insurance (Tselentis *et al.*, 2016) includes *pay as you drive* and *pay how you drive* schemes. The pay as you drive scheme focuses on the driving habits, e.g. the driven distance, the time of day and how long the insured person has been driving. Pay how you drive also considers the driving style, e.g. the speed, harsh or smooth braking, aggressive acceleration or deceleration, cornering and parking skills.

Telematics insurance started as a niche market when the technology first surfaced more than 10 years ago. Early adopters of UBI were seen in the USA, Italy and the UK. On April 28th, 2015, the European Parliament voted in favour of eCall regulation which forces all new cars in the EU from April 2018 onwards to be equipped with a telematics device that will automatically dial 112 in the event of an accident, providing precise location and impact data (Regulation (EU) 2015/758 of the European Parliament and of the Council of April 29th, 2015, concerning type approval requirements for the deployment of the eCall in-vehicle system based on the 112 service and amending Directive 2007/46/EC).

These potentially high dimensional telematics data, collected on the fly, force pricing actuaries to change their current practice, both from a business as well as a statistical point of view. New statistical models must be developed to set premiums adequately on the basis of an individual policyholder's driving habits and style and the current literature on insurance rating does not adequately address this question. In this paper, we take a first step in this direction. We use a Belgian telematics insurance data set with in total over 297 million km driven. On the basis of how many kilometres the insured person drivers, on which kind of roads and during which moments in the day, we quantify the effect of individual driving habits on expected claim frequencies. Combined with a similar predictive model for claim severities, which is outside the scope in this paper, this enables tailor-made car insurance pricing. We first discuss how a car insurance policy is traditionally priced and relate this to the literature investigating the effect of vehicle usage on accident risk in Section 2. The data set is described in Section 3, along with the necessary preliminary data processing steps to combine the telematics information with the policy and claims information. By constructing predictive models for claim frequency, we compare the performance of different sets of predictor variables (e.g. traditional *versus* purely telematics) and unravel the relevance and influence of adding telematics insights. In particular, we contrast the use of time and distance as exposure-to-risk measures. The novel methodological contribution of this paper (Section 4), incorporates the divisions of the driven distance by road type and time slots as compositional predictors in the regression framework of generalized additive models (GAMs) and constructs a new way to interpret and visualize their effect on the average claim frequency. We develop both a conditioning and a projection approach to handle structural 0s in one or more components of a compositional predictor. We present the results in Section 5 whereas Section 6 concludes.

Accompanying R code which illustrates how to apply the methods on data with a similar structure can be obtained from

```
http://wileyonlinelibrary.com/journal/rss-datasets
```

2. Statistical background and related modelling literature

Insurance pricing is the calculation of a fair premium, given the policy(holder) characteristics, as well as information on claims reported in the past (if available). Pricing relies on regression techniques and requires a data set with policy(holder) information and corresponding claim frequencies and severities, where severity is the ultimate total impact of a claim. The claim frequency and severity components are typically modelled separately by using regression techniques (Frees, 2014). The current state of the art (see Denuit *et al.* (2007) and de Jong and Heller (2008) for an overview) uses generalized linear models (GLMs) (McCullagh and Nelder, 1989), with typically a Poisson GLM for the claim counts and a gamma GLM for the claim severities. In car insurance, the duration of the policy period during which coverage is provided is referred to as the *exposure to risk*. The expected number of claims is in practice modelled directly proportional to the exposure, to make the premiums proportional to the length of coverage. From a theoretical point of view, this can be motivated by the probabilistic framework of Poisson processes (Denuit *et al.*, 2007). It is, however, suggested (see for example Butler (1993)) that every kilometre travelled by a vehicle transfers risk to its insurer and hence the number of driven kilometres (*car-kilometres*) should be adopted as the exposure unit instead of the policy duration (*car-years*). Statistical studies show how claim frequencies significantly increase with kilometres (Bordoff and Noel, 2008; Ferreira and Minikel, 2010; Litman, 2011; Boucher *et al.*, 2013; Lemaire *et al.*, 2016). Most of these studies show a relationship between claim frequencies and the number of driven kilometres which is less than proportional. One of the focus points in our study (Section 3.2) is to investigate the relationship between the expected number of claims and both exposure-to-risk measures (i.e. time and distance).

Using models involving both policy and telematics predictors, Ayuso *et al.* (2014, 2016a) studied the travelled time and distance to the first accident by using Weibull regression models. Paefgen *et al.* (2014) investigated the relationship between accident risk and driving habits by using logistic regression models. Their case–control study design does not allow for inference on the probability of accident involvement. The difference in time exposure between the vehicles with accident involvement (6 months before the accident) and the control group (24 months) was, however, only used to obtain a per-month distance exposure but was further neglected in the study. Traditional risk factors were not accounted for, since that information was not available, and the compositional nature of the telematics predictor variables constructed was ignored. In contrast (Section 3.2), the main focus points in our research are

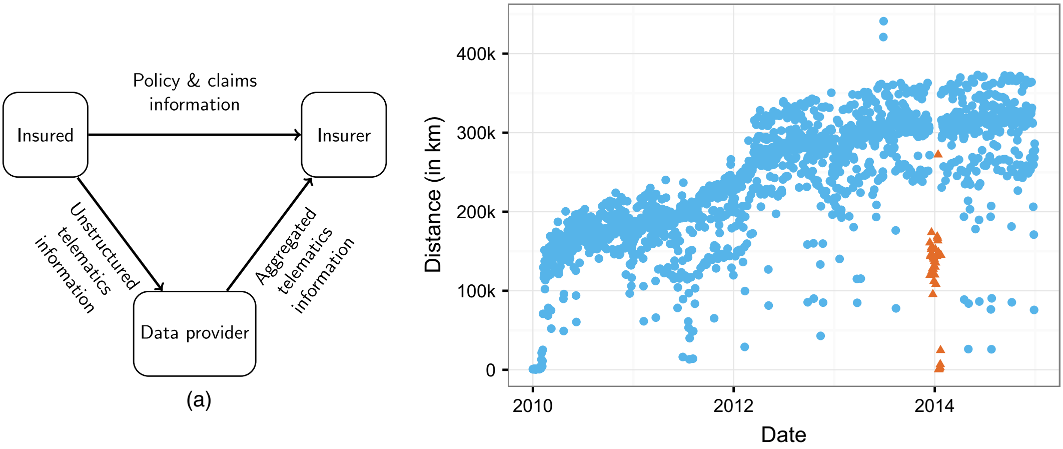
(a) combining the new telematics variables with traditional policy(holder) information through a careful model and variable selection process and
(b) incorporating the compositional structure of the telematics variables in the analysis.

3. Telematics insurance data

We consider data from a Belgian portfolio of drivers with motor third-party liability (MTPL) insurance. MTPL insurance is the legally compulsory minimum insurance covering damage to third parties' health and property caused by an accident for which the driver of the vehicle is responsible. The special type of MTPL product that we are considering is specifically aiming for young drivers who traditionally face high insurance premiums. Insured drivers were offered a substantial discount on their premium if they agreed to install a telematics black box device in their car. The telematics box collects statistics on driving habits: how often one drivers, how many kilometres, where and when. Information on the driving style (such as speeding, braking, accelerating, cornering or parking) is not registered. The telematics data did not have an effect on the (future) premium levels of the drivers and did not induce any restrictions on how much or where they can drive.

3.1. Data processing

The unstructured telematics data, which were collected by the telematics box that was installed in the vehicle, are first transmitted to the data provider who structures and aggregates these data each day and then reports them to the insurance company as a comma-separated-values file (Fig. 1(a)). Only the structured, aggregated telematics information is available to us. Each daily file contains information on the daily driven distance (in metres) for each policyholder. This number of metres is split into four road types (*urban*, *other*, *motorways* and *abroad*) and five time slots (*6h-9h30*, *9h30-16h*, *16h-19h*, *19h-22h* and *22h-6h*). The nature of the data does



(a)

(b)

Fig. 1. (a) Schematic overview of the flow of information and (b) number of registered kilometres on each day on an aggregate, portfolio level for the telematics data observed between January 1st, 2010, and December 31st, 2014: ▲ outliers by the turn of the year 2014, corresponding to a technical malfunction not allow for a classification of a driven metre by road type and time slot simultaneously. The number of trips, measured as key-on–key-off events, is also reported. This is a typical set-up (see Paefgen *et al.* (2014)). In this study, we analyse the telematics data that were collected between January 1st, 2010, and December 31st, 2014.

The telematics data are linked to the policy(holder) and claims information of the insurance company corresponding to the portfolio under consideration (see Table 1 for a complete list). Policy data, such as age, gender and characteristics of the car, are directly reported by the insured driver to the insurer at underwriting (see Fig. 1(a)). They are updated over time which enables us to link the claims occurring at a specific moment in time to the correct policy information. Each observation of a policyholder in the policy data set refers to a policy period over which the MTPL insurance coverage holds and contains the most recent policy information. For most insured drivers, this coverage period is 1 year; however, it can be smaller for several reasons. If for instance the policyholder decides to add comprehensive coverage, buys a new vehicle or changes his residence during the term of the contract, the policy period will be restricted to the date of the policy modification and an additional observation line will be added for the subsequent period. A policy period can also be split when the coverage is suspended for a certain time.

Using the policy number and period we first merge the telematics information at daily level with the policy data set. Next, we adjust the start and end date of the policy periods on the

Table 1. Description of the variables contained in the data set arising from the various sources of information

Variable	*Description*
Claims information	
claims	Number of reported MTPL claims at fault during the policy period
Policy information	
policy period	Duration in days of the policy period (minimal 30 days and at most 1 year)
age	Age of the least experienced driver listed on the policy at the start of the policy period, measured as the number of years between the date of birth and the start of the policy period
experience	Experience of the least experienced driver listed on the policy, measured as the number of years between the date when the driver's licence was obtained and the start of the policy period
gender	Gender of the least experienced driver listed on the policy (*male* or *female*)
material damage cover	Indicator whether the insurance policy also covers material damage (*yes* or *no*)
postal code	Belgian postal code where the policyholder resides
bonus-malus	*Bonus–malus* level of the policy, reflecting the past individual claims experience, between −4 and 22 with lower values indicating a better history
age vehicle	Age of the vehicle, measured as the number of years between the date when the car was registered and the start of the policy period
kwatt	Horsepower of the vehicle, measured in kilowatts
fuel	Fuel type of the vehicle (*petrol* or *diesel*)
Telematics information	
distance	Distance in metres driven during the policy period
yearly distance	Distance in metres driven during the policy period, rescaled to a full year by dividing by the duration in days of the policy period and multiplying by 365
trips	Number of trips (*key-on*, *key-off*) during the policy period
average distance	Distance in metres driven on average during one trip, obtained by dividing the distance by the number of trips
road type	Division of the distance into 4 road types (*urban*, *other*, *motorways* and *abroad*)
time slot	Division of the distance into 5 time slots (*6h-9h30*, *9h30-16h*, *16h-19h*, *19h-22h* and *22h-6h*)
week/weekend	Division of distance into *week* (Monday–Friday) and *weekend* (Saturday and Sunday)

basis of the first and last day at which telematics data are observed for each policy period of each insured person. This ensures that the adjusted policy periods reflect time periods over which both the insurance coverage holds and telematics data are collected. On the basis of Fig. 1(b), where we plot the evolution of the driven distance on each day by all drivers of the portfolio, we suspect that technical deficiencies of the data provider can cause under-reporting of the number of metres driven at an aggregate level. The outliers indicated as triangles by the turn of the year 2014 could be linked to a serious technical failure preventing telematics information from being reported for a significant part of our portfolio. We dealt with this by removing this period of roughly 1 month from the policy periods of all the insured drivers. In the remainder of the observation period between January 1st, 2010, and December 31st, 2014, clear causes of under-reporting could not be identified and hence we did not take any other corrective action. However, this illustrates that data reliability forms a challenge for this new telematics technology. We further removed those observations with a policy duration of less than 30 days to avoid senseless observations of only a couple of days and retained only the complete observations with no missing policyholder information.

Next, we aggregate the telematics information by policyholder and period. This means that we sum the distance driven, their divisions into four road types and five time slots, and the number of trips made. Finally, we use the claims information to extract the number of MTPL claims at fault that occurred between the start and end date of the adjusted policy periods for each policy record.

Over the time period of this study, we end up with a data set of 33 259 observations. Table 1 gives an overview of the available variables coming from the three sources of data (claims, policy and telematics). These observations correspond to 10 406 unique policyholders, who are followed over time, have jointly driven over 297 million km during a combined insured policy period of 17 681 years and reported 1481 MTPL claims at fault. Hence, on average, there were 0.0838 claims per insured year or 0.0499 claims per 10 000 driven km. For over 95% of the observations no claim occurred during the corresponding policy period, whereas for 52 observations two claims occurred and for a single observation even three during the same policy period.

3.2. Risk classification using policy and telematics information

The goal of this research is to build a rating model to express the number of claims as a function of the available covariates. Two sources of information are combined which are described in detail in Table 1. First, there is the self-reported policy information which contains all rating variables that are traditionally used in car insurance pricing. The second source of information is derived from the telematics data. The main objective is to discover the relevance and influence of adding the new telematics insights by using flexible statistical modelling techniques in combination with appropriate model and variable selection tools. One of the key questions is whether the risk that is transferred from the policyholder to the insurer is proportional to the duration of the policy period or the driven distance during that time. Telematics technology allows a shift to be made from time as exposure to distance as exposure. This would lead to a form of pay as you drive insurance, where a driver pays for every kilometre driven. Histograms of both potential exposure variables are contrasted in Figs 2(a) and 2(b).

To investigate the influence and explanatory power of the telematics variables in predicting the risk of an accident, we compare the performance of four sets of predictor variables used to model the number of claims; see Fig. 2(c). The *classical* set contains only policy information and uses time as exposure to risk. The *telematics* set contains only telematics information and uses the distance in metres as exposure to risk. The two other models, *time hybrid* and *metre hybrid*,

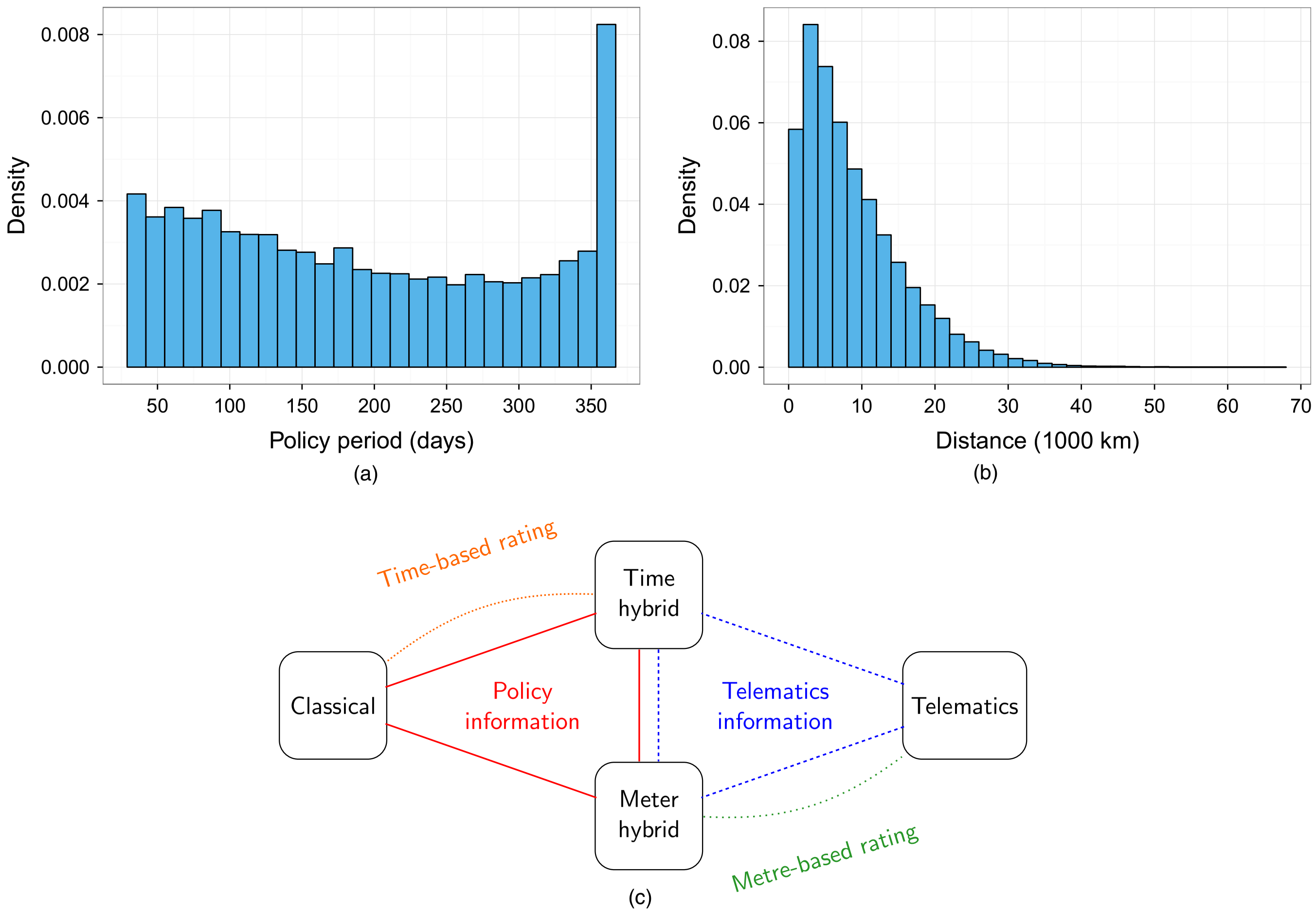


Fig. 2. Histogram of (a) the duration (in days) of the policy period (at most 1 year) and (b) the distance driven (in 1000 kilometres) during the policy period, and (c) graphical representation of the similarities and differences between the four predictor sets

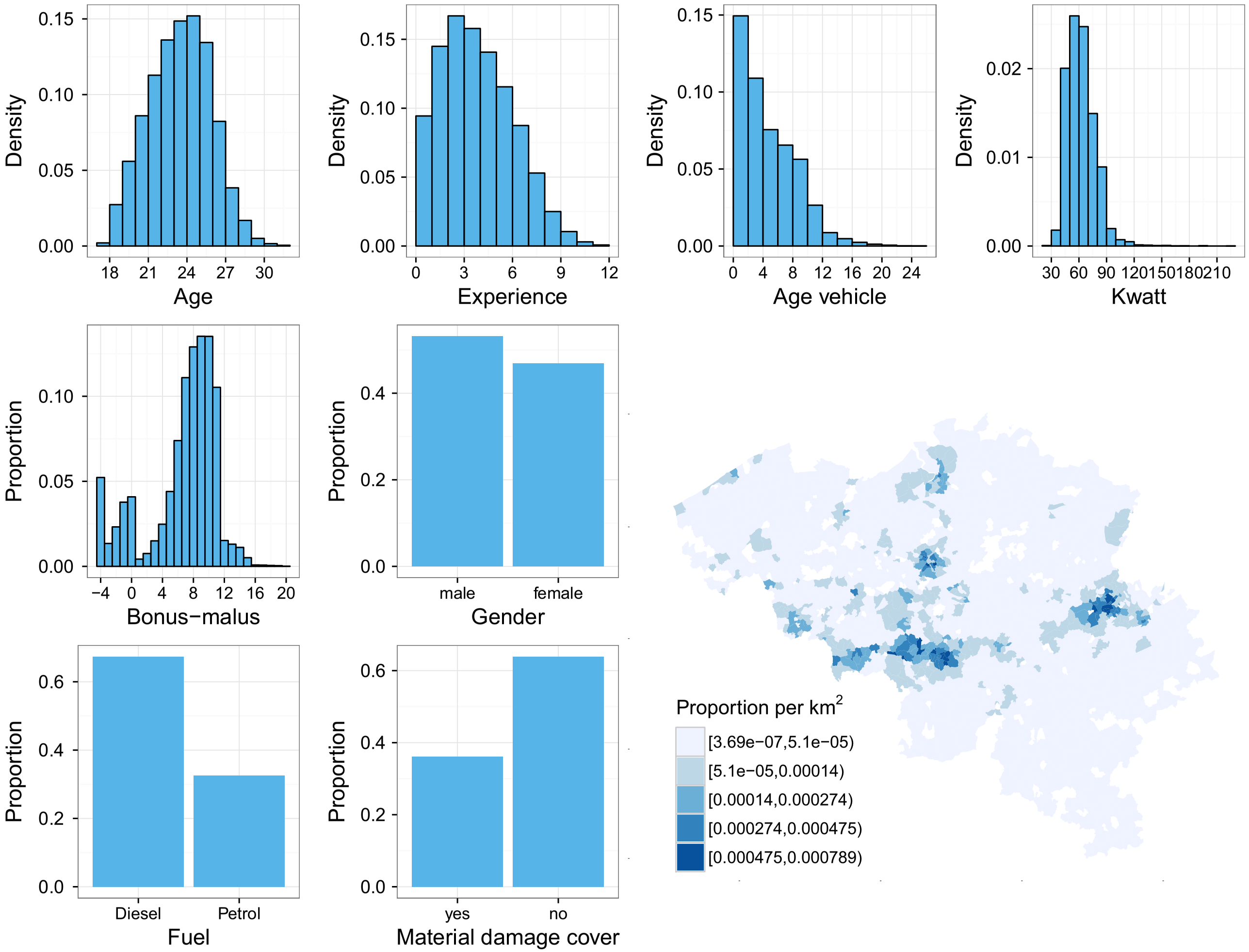


Fig. 3. Histograms and bar plots of the continuous and categorical policy variables contained in the data set: the map depicts the geographical information by showing the proportion of insureds per squared kilometre living in each of the different postal codes in Belgium; the five class intervals have been created by using *k*-means clustering

both contain policy and telematics information. Whereas the first uses time as an exposure measure, the second uses distance. These four predictor sets contrast on the one hand the use of traditional policy rating variables and telematics variables and on the other hand the use of policy duration *versus* distance as exposure measures in the assessment of the risk.

The main predictors based on the policy information besides the duration of the policy period include the age of the driver, the experience as measured using the driver's licence age, the gender, characteristics of the car and the postal code where the policyholder lives. In the case of multiple insured drivers (around 18% of the observations), we select (in consultation with the insurer) the age, gender, experience and postal code belonging to the driver with the most recent permit and hence the lowest experience. This is in line with the strategy of the insurer who offers this type of insurance contract to young drivers. The *bonus–malus* (BM) level is a special kind of variable that reflects the past individual claims experience. It is a function of the number of claims reported in previous years with values between −4 and 22 where lower levels indicate a better history. The insurer uses a slightly modified version of the former compulsory Belgian BM system, which all companies operating in Belgium have been obliged to use from 1992 to 2002, with minor refinements for the policyholders occupying the lowest levels in the scale. Despite the deregulation, many insurers in the Belgian market still apply the former mandatory system (Denuit *et al.*, 2007). Even though the BM scale level is not a covariate of the same type as the other *a priori* variables, we keep it in the analysis to have an idea of the information that is contained in this variable (as was also done in, for instance, Denuit and Lang (2004)). From a statistical point of view, it tries to structure dependences between observations arising from the same policyholder. An overview of the policy predictor variables and their sample distributions is given in Fig. 3.

In the telematics information set we use the distance driven during the policy period as a predictor but we also create two additional telematics variables: the yearly and average distance driven; see Table 1. Histograms of these variables are shown in Fig. 4. The divisions of the driven distance by time slot, road type and week or weekend are highly correlated with the total distance driven as they sum to this amount. To distinguish the absolute information that is measured by the driven distance in a certain policy period from the compositional information of the distance split into different categories, we consider boxplots of the relative proportions in Fig. 4. These relative proportions sum to 1 for each observation. To stress this interconnectedness in the different splits, we show the compositional profiles of a sample of 100 drivers on top of the marginal boxplots. Another important point to stress is that not all components of a certain division of the distance are present for each observation. For instance, if an insured person does not drive abroad during the policy period, the relative proportion of the distance driven abroad will be 0. The use of such compositional information as predictors in statistical modelling is a key issue in this research.

4. Model building and selection

We model the frequencies of claims by constructing Poisson and negative binomial (NB) regression models. We denote by N_{it} the number of claims for policyholder i in policy period t with $i=1,\ldots,I$ and $t=1,\ldots,T_i$. The model is denoted by $N_{it}\sim\text{Poisson}(\mu_{it})$ or $N_{it}\sim\text{NB}(\mu_{it},\phi)$, where $\mu_{it}=\mathbb{E}(N_{it})$ represents the expected number of claims reported by policyholder i in policy period t and ϕ is the parameter of the NB distribution such that $\text{var}(N_{it})=\mu_{it}+\mu_{it}^2/\phi$, allowing for overdisperion. A log-linear relationship between the mean and the predictor variables is specified by the log-link function. This means that we set $\mu_{it}=\exp(\eta_{it})$ where η_{it} is a predictor function of the available explanatory factors. The probability mass functions for the Poisson and the NB models are respectively expressed as

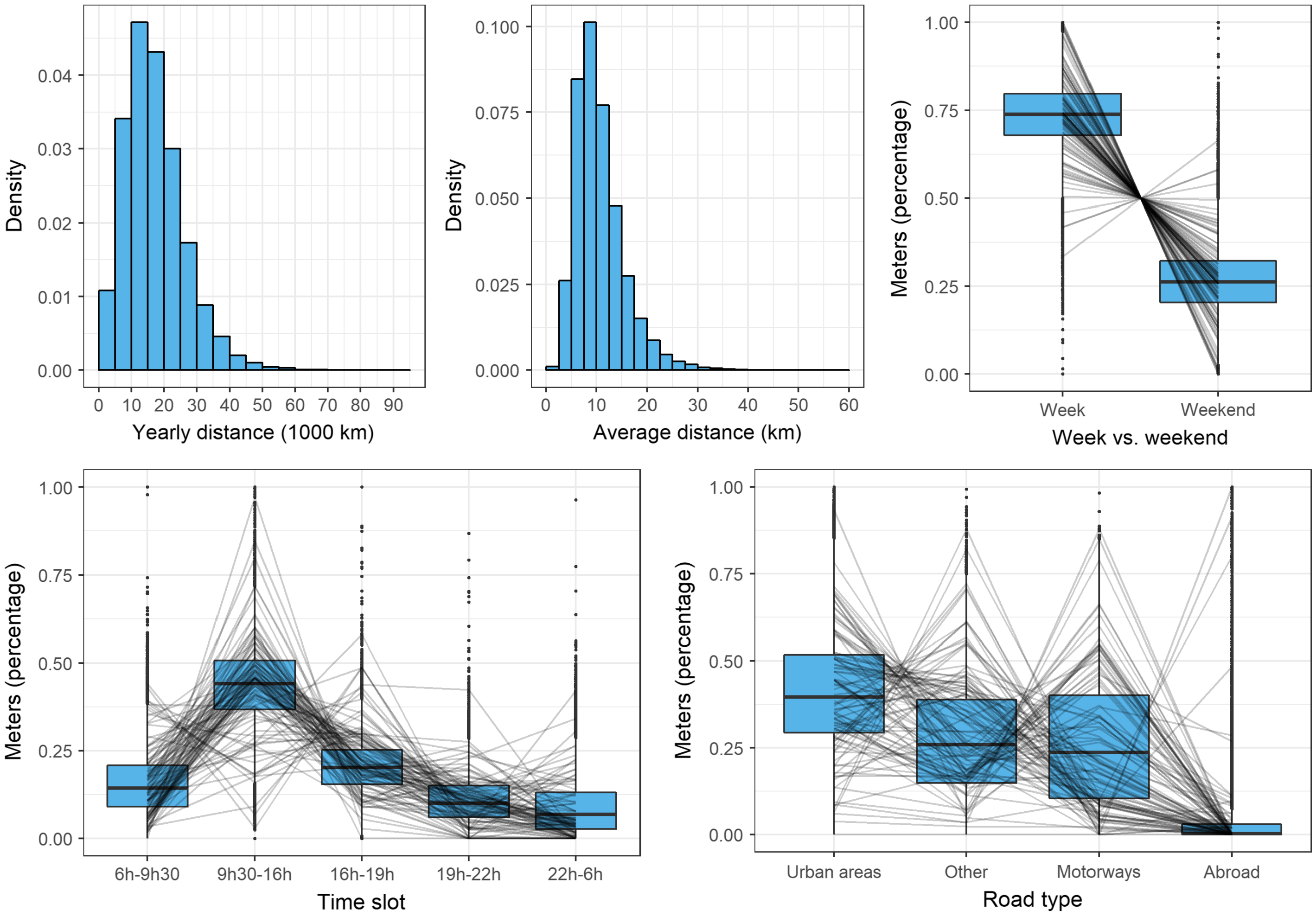


Fig. 4. Graphical illustration of the telematics variables contained in the data set: for the yearly and average distance, we construct histograms; for the division of the driven distance by road types, time slots and week or weekend, we construct boxplots of the relative proportions; to highlight the dependences intrinsic to the fact that the division into different categories sums to 1, we plot profile lines for 100 randomly selected observations in the data set

$$\mathbb{P}(N_{it}=n_{it})=\frac{\exp(-\mu_{it})\mu_{it}^{n_{it}}}{n_{it}!}$$

and

$$\mathbb{P}(N_{it}=n_{it})=\left(\frac{\phi}{\phi+\mu_{it}}\right)^{\phi}\frac{\Gamma(\phi+n_{it})}{n_{it}!\,\Gamma(\phi)}\left(\frac{\mu_{it}}{\phi+\mu_{it}}\right)^{n_{it}}.$$

For each of the predictor sets in Fig. 2(c) we construct the best model by using the allowed information based on the Akaike information criterion AIC; see Section 4.3. Additionally, we identify the best models under the restriction that the risk is proportional to the time or metre exposure. This is accomplished by incorporating the logarithm of the exposure to risk, either duration of the policy period or total distance driven during the policy period, as an offset term in the predictor, i.e. a regression variable with a constant coefficient of 1 for each observation. In the most general case, the predictor has the form

$$\eta_{it}=\beta_0+\text{offset}+\eta_{it}^{\text{cat}}+\eta_{it}^{\text{cont}}+\eta_{it}^{\text{spatial}}+\eta_i^{\text{re}}+\eta_{it}^{\text{comp}}, \tag{1}$$

where β_0 denotes the intercept, the categorical effects are bundled in η_{it}^{cat}, the term η_{it}^{cont} contains the effects of the continuous predictors, $\eta_{it}^{\text{spatial}}$ represents the geographical effect, η_i^{re} the policyholder-specific random effect and the term η_{it}^{comp} embodies the effects of the compositional predictors. Under the offset restriction, the continuous effect of the exposure to risk, either the duration of the policy period (time-based rating) or the distance driven (metre-based rating), is replaced by the logarithm of the exposure to risk as an offset.

Zero-inflated variants of these models are not considered because of interpretability reasons. Such models cannot capture the effect of a varying exposure to risk in a transparent and intuitive way.

4.1. Generalized additive models

The model framework that we work that with in this study is that of GAMs, introduced by Hastie and Tibshirani (1986). GAMs enable us to incorporate continuous covariates in a more flexible way compared with the traditional GLMs that are used in actuarial practice (see for example Klein *et al.* (2014)). From a standpoint of accuracy, GAMs are competitive with popular black box machine learning techniques (such as neural networks, random forests or support vector machines), but they have the important advantage of interpretability. In insurance pricing it is of crucial importance to have interpretable results to understand the premium structure and to explain this to clients and regulators. Using a semiparametric additive structure, GAMs define non-parametric relationships between the response and the continuous variables in the predictor in the following way:

$$\eta_{it}^{\text{cat}}+\eta_{it}^{\text{cont}}=\mathbf{Z}_{it}\boldsymbol{\beta}+\sum_{j=1}^{J}f_j(x_{jit}),$$

where $\mathbf{Z}_{it}$ represents the row corresponding to policyholder i in policy period t of the model matrix of the categorical variables with parameter vector $\boldsymbol{\beta}$ and f_j represents a smooth function of the jth continuous predictor variable. To estimate f_j, we choose cubic spline basis functions B_{jk}, such that in our models $f_j(x)=\Sigma_{k=1}^{q}\gamma_{jk}B_{jk}(x)$. The knots are chosen by using 10 quantiles of the unique x_j-values. Cardinal basis functions parameterize the spline in terms of its values at the knots (Lancaster and Salkauskas, 1986). For identifiability, we impose constraints by centring each smooth component around zero; thus $\Sigma_{i=1}^{I}\Sigma_{t=1}^{T_i}f_j(x_{jit})=0$ for $j=1,\ldots,J$. To avoid overfitting, the cubic splines are penalized by the integrated squared second derivative (Green and Silverman, 1994), which yields a measure for the overall curvature of the function. For each component, this penalty can be written as a quadratic function:

$$\int f_j''(x)^2 \,\mathrm{d}x = \sum_{k=1}^{q} \sum_{l=1}^{q} \gamma_{jk}\gamma_{jl} \int B_{jk}''(x) B_{jl}''(x) \,\mathrm{d}x = \boldsymbol{\gamma}_j^{\mathrm{T}} \mathbf{S}_j \boldsymbol{\gamma}_j,$$

with $(\mathbf{S}_j)_{kl} = \int B_{jk}''(x) B_{jl}''(x) \mathrm{d}x$. Given these penalty functions for each component, we define the penalized log-likelihood as

$$l(\boldsymbol{\psi}) - \frac{1}{2} \sum_{j=1}^{J} \lambda_j \boldsymbol{\gamma}_j^{\mathrm{T}} \mathbf{S}_j \boldsymbol{\gamma}_j, \tag{2}$$

where $l(\boldsymbol{\psi})$ denotes the log-likelihood as a function of all model parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J)^{\mathrm{T}}$ and λ_j denotes the smoothness parameter that controls the trade-off between goodness of fit and the degree of smoothness of component f_j for $j = 1, \ldots, J$. Different smoothing parameters for each component enable us to penalize the smooth functions differently.

The model parameters $\boldsymbol{\psi}$ are estimated by maximizing expression (2) by using penalized iteratively reweighted least squares (Wood, 2006). For the Poisson model, the smoothing parameters $\lambda_1, \ldots, \lambda_J$ are estimated by using an unbiased risk estimator criterion, which is a rescaled version of Akaike's information criterion AIC (Akaike, 1974). For the NB model, we estimate the smoothing parameters and the scale parameter ϕ by using maximum likelihood. Alternatively, the smoothing parameters can also be estimated by using restricted maximum likelihood (Krivobokova and Kauermann, 2007; Reiss and Ogden, 2009; Wood, 2011).

In addition to categorical and continuous covariates, the data set contains spatial information, namely the postal code where the policyholder resides. Insurance companies tend to use the geographical information of the insured person's residence as a proxy for the traffic density and for other unobserved sociodemographic factors of the neighbourhood. We model the spatial heterogeneity of claim frequencies by adding a spatial term $\eta_{it}^{\mathrm{spatial}} = f_s(\mathrm{lat}_{\mathrm{it}}, \mathrm{long}_{it})$ in the additive predictor η_{it}, using the latitude and longitude co-ordinates (in degrees) of the centre of the postal code where the policyholder resides. We use second-order smoothing splines on the sphere (Wahba, 1981) to model f_s. This enables us to quantify the effect of the geographic location while taking the regional closeness of the neighbouring postal codes into account.

In our data set, many policyholders $i = 1, \ldots, I$ are observed over multiple policy periods $t = 1, \ldots, T_i$. This longitudinal aspect of the data can be modelled by including policyholder-specific random effects η_i^{re} in the predictor. The GAM considered thus far is extended in this way by exploiting the link between penalized estimation and random effects (see for example Ruppert *et al.* (2003)). We assess whether such random effects are needed to take the correlations between observations of the same policyholder into account by using the approximate test for a zero random effect that was developed by Wood (2013).

4.2. Compositional data

The divisions of the total driven distance into the various categories—road types (four), time slots (five) and week or weekend (two) (see Table 1)—are highly correlated with and sum to the total distance driven. Hence in Fig. 4 we divided all components of each split by the total distance driven. Incorporating these divisions, either in absolute or relative terms, in a predictor also containing the total distance leads to a perfect multicollinearity problem. The most straightforward way to deal with this would be to leave one component out, but this approach is not permutation invariant and the statistical inference will depend on which component is removed, making interpretations misleading. The standard regression interpretation of a change in one of the components of the distance when the other components are held constant is not possible because of the sum constraint. We introduce and further develop the necessary statistical tools to model such predictors.

In the literature, data which quantitatively describe the parts of some whole and provide only relative information between their components are called *compositional data* (van den Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn *et al.*, 2015). Typical examples include mineral compositions, molar concentrations and household budgets. In our setting, the divisions of the distance driven are compositional. Scale invariance is a key property: if a composition is scaled by a constant, the information that is carried is completely equivalent. Therefore compositional data can be represented by real vectors with positive components that sum to 1. The space of representations of compositions is called the simplex of D parts, defined by

$$\mathcal{S}^D = \{\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}} : x_i > 0, \sum_{i=1}^{D} x_i = 1\}.$$

When data are considered compositional, classical statistics, that do not take the special geometry of the simplex into account, are not appropriate. Section 4.2.1 revises the necessary geometrical concepts to work with compositional data. Extending the current literature, we propose a new way of quantifying and interpreting the effect of the compositional explanatory variables on the outcome in Section 4.2.2. Section 4.2.3 introduces two approaches to accommodate structural 0s in regression with compositional predictors.

4.2.1. The Aitchison geometry of the simplex

Aitchison (1986) introduced operations between compositional data vectors which define a vector space structure on the mathematical simplex known as the *Aitchison geometry of the simplex*. *Perturbation* plays the role of addition on the simplex and is defined as a closed componentwise product $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D)^{\mathrm{T}}$, where the closing operation $\mathcal{C}$ ensures a total sum of 1, i.e. the closure of $\mathbf{x}$ is $\mathcal{C}(\mathbf{x}) = \mathbf{x}/\Sigma_{i=1}^{D} x_i$. The product of a vector by a scalar is called *powering* and is defined as the closed componentwise powering of a composition by a scalar, i.e. $\alpha \odot \mathbf{x} = \mathcal{C}(x_1^{\alpha}, \ldots, x_D^{\alpha})^{\mathrm{T}}$, for $\alpha \in \mathbb{R}$. The *Aitchison inner product* for compositions,

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathrm{a}} = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln\left(\frac{x_i}{x_j}\right) \ln\left(\frac{y_i}{y_j}\right) = \sum_{i=1}^{D} \ln(x_i) \ln(y_i) - \frac{1}{D} \left\{ \sum_{i=1}^{D} \ln(x_i) \right\} \sum_{j=1}^{D} \ln(y_j)$$

is proportional to the scalar product of the vectors that is formed by all possible pairwise log-ratios of the two compositions and induces the norm $\|\mathbf{x}\|_{\mathrm{a}} = \surd \langle \mathbf{x}, \mathbf{x} \rangle_{\mathrm{a}}$ and distance $d_{\mathrm{a}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_{\mathrm{a}}$, where '$\ominus$' represents the opposite operation of '$\oplus$', i.e. $\ominus \mathbf{y} = \oplus\{(-1) \odot \mathbf{y}\}$, a closed componentwise division. The simplex along with these operations then forms a $(D-1)$-dimensional Euclidean vector space $(\mathcal{S}^D, \oplus, \odot, \langle \cdot, \cdot \rangle_{\mathrm{a}})$. Given this Euclidean structure, we can measure distances and angles, and define related geometrical concepts. Elementary statistical notions involving the metrics of the sample space can be adapted to the Euclidean structure of the simplex.

Compositional data are analysed by using a log-ratio approach and compositions $\mathbf{x}$ from the simplex can be represented in the real space by using the *centred log-ratio transformation clr*:

$$u_i = \mathrm{clr}_i(\mathbf{x}) = \ln\left\{ \frac{x_i}{g(\mathbf{x})} \right\}, \qquad i = 1, \ldots, D, \quad g(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i \right)^{1/D} \tag{3}$$

where $g(\mathbf{x})$ denotes the geometric mean of the components. The clr-transformation defines an isometry between $\mathcal{S}^D$ and the $(D-1)$-dimensional subspace of $\mathbb{R}^D$ of vectors whose components add to 0, denoted by $\mathcal{H}^D = \{\mathbf{u} \in \mathbb{R}^D | \Sigma_{i=1}^D u_i = 0\}$ and called the clr-plane. Using matrix notation we can write the clr-transform and its inverse as

$$\begin{aligned} \mathbf{u} &= \mathrm{clr}(\mathbf{x}) = \ln\{\mathbf{x}/g(\mathbf{x})\}, \\ \mathbf{x} &= \mathrm{clr}^{-1}(\mathbf{u}) = \mathcal{C}\{\exp(\mathbf{u})\}, \end{aligned} \tag{4}$$

where the logarithmic and exponential function apply componentwise. An orthonormal basis of $\mathcal{S}^D$ can be obtained from an orthonormal basis of $\mathcal{H}^D$ by using the inverse clr-transformation. A transformation between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$ that provides the co-ordinates of any composition with respect to a given orthonormal basis is called an *isometric log-ratio transformation ilr*. The transformation that was originally defined by Egozcue *et al.* (2003) maps a compositional data vector $\mathbf{x}$ in a $(D-1)$-dimensional real vector $\mathbf{z} = (z_1, z_2, \ldots, z_{D-1})^{\mathrm{T}}$ with components

$$z_i = \mathrm{ilr}_i(\mathbf{x}) = \sqrt{\left(\frac{D-i}{D-i+1}\right)} \ln\left\{ \frac{x_i}{\sqrt[D-i]{\left(\prod_{j=i+1}^{D} x_j\right)}} \right\}, \qquad i = 1, \ldots, D-1. \tag{5}$$

By arranging the corresponding orthonormal basis vectors in $\mathcal{H}^D$ by columns, we obtain a $D \times (D-1)$ matrix V with elements

$$V_{ij} = \begin{cases} \dfrac{D-j}{\surd\{(D-j+1)(D-j)\}} & \text{for } i = j, \\ \dfrac{-1}{\surd\{(D-j+1)(D-j)\}} & \text{for } i > j \end{cases}$$

and 0 otherwise, for which it holds that $V^{\mathrm{T}}V = I_{D-1}$ and $VV^{\mathrm{T}} = I_D - (1/D)\mathbf{1}_D\mathbf{1}_D^{\mathrm{T}}$, where I_D is the identity matrix of dimension D and $\mathbf{1}_D$ is a D-vector of 1s (Egozcue *et al.*, 2011). Then we can write the ilr-transform and its inverse as

$$\begin{aligned} \mathbf{z} &= \mathrm{ilr}(\mathbf{x}) = V^{\mathrm{T}} \ln(\mathbf{x}) = V^{\mathrm{T}} \mathrm{clr}(\mathbf{x}), \\ \mathbf{x} &= \mathrm{ilr}^{-1}(\mathbf{z}) = \mathcal{C}\{\exp(V\mathbf{z})\}. \end{aligned} \tag{6}$$

The clr- and ilr-transformations reflect how all relevant information of a composition is conveyed by the component log-ratios. In the case $D = 2$, the ilr-transformation (5) is proportional to the logit function, which is used in logistic regression to transform the probability $0 < p < 1$ of a binary response into unrestricted log-odds.

As the clr- and ilr-transformations are isometric, all angles and distances are preserved. This means that, whenever compositions are transformed into co-ordinates, the metrics and operations in the Aitchison geometry of the simplex are translated into the ordinary Euclidean metrics and operations in real space. For instance, the Aitchison inner product of two compositions is equal to the real inner product of their clr- or ilr-transformed vectors:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathrm{a}} = \langle \mathrm{clr}(\mathbf{x}),\ \mathrm{clr}(\mathbf{y}) \rangle = \sum_{i=1}^{D} \mathrm{clr}_i(\mathbf{x})\, \mathrm{clr}_i(\mathbf{y}) = \langle\, \mathrm{ilr}(\mathbf{x}),\ \mathrm{ilr}(\mathbf{y}) \rangle = \sum_{i=1}^{D-1} \mathrm{ilr}_i(\mathbf{x})\, \mathrm{ilr}_i(\mathbf{y}).$$

Even though the simplex $\mathcal{S}^D$ is a subset of the real space $\mathbb{R}^D$, Aitchison (1986) showed that the geometry is clearly different. Ignoring the compositional nature of the data in a statistical context can lead to incompatible or incoherent results. The principle of working on co-ordinates (Mateu-

Figueras *et al.*, 2011) is first to express the compositional data with respect to an orthonormal basis of the underlying vector space with Euclidean structure, next, to apply standard statistical techniques to the vectors of co-ordinates and, finally, to back-transform and describe the results in terms of the simplex. Final results do not depend on the basis chosen.

4.2.2. A new interpretation for compositional predictors

In our framework, the total distance in metres is used as a continuous predictor in the telematics models and its effect is modelled by using a smooth function. In addition, we propose to treat the divisions of the driven distance by road types, time slots and week or weekend as compositional data covariates in the claim count regression models. In this way, the effects of the absolute information of the total distance driven and the relative information that is contained in the different divisions can be structured and interpreted separately.

In the context of linear regression, Hron *et al.* (2012) proposed first to apply the isometric log-ratio transform (5) to map the compositions in the D-part Aitchison simplex to a $D-1$ Euclidean space before including them as explanatory variables. More generally, in any regression context involving a predictor, we can add a *compositional predictor* term η^{comp} by using the ilr-transformed variables, i.e.

$$\eta^{\text{comp}} = \beta_1 z_1 + \ldots + \beta_{D-1} z_{D-1}. \tag{7}$$

The model fitted does not depend on the choice of the orthonormal ilr-basis since the co-ordinates of $\mathbf{x}$ with respect to different orthonormal bases are orthogonal transformations of each other. Using the ilr-transformation the model parameters can be estimated without constraints and the *ceteris paribus* interpretation of altering one z_i without altering any other becomes possible. A drawback is that only the first regression parameter, β_1, has a comprehensible interpretation since z_1 explains relevant information about x_1. The remaining coefficients are not straightforward to interpret. Also the suggestion by Hron *et al.* (2012) to permute the indices in formula (5) and to construct D regression models, each time with a different component first, is undesirable. This is especially so in our case where we have more than one compositional predictor and each model fit is computationally intensive because of smooth continuous, spatial and random effects.

Hence, we develop a new way to interpret and visualize the effect of a compositional predictor (7) without the need to refit the model. Following van den Boogaart and Tolosana-Delgado (2013) and Pawlowsky-Glahn *et al.* (2015), we start by using the inverse ilr-transform on the model coefficients, i.e. set $\mathbf{b} = \text{ilr}^{-1}(\boldsymbol{\beta})$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{D-1})^{\text{T}}$, such that we can rewrite the compositional predictor as

$$\eta^{\text{comp}} = \sum_{i=1}^{D-1} \beta_i z_i = \sum_{i=1}^{D-1} \text{ilr}_i(\mathbf{b})\ \text{ilr}_i(\mathbf{x}) = \langle \mathbf{b}, \mathbf{x} \rangle_{\text{a}}.$$

The composition $\mathbf{b} \in \mathcal{S}^D$ can be interpreted as the simplicial gradient of η^{comp} with respect to $\mathbf{x}$ (Barceló-Vidal *et al.*, 2011) and is the compositional direction along which the predictor increases fastest. In particular, if we perturb $\mathbf{x}$ by a unit vector $\mathbf{b}/\|\mathbf{b}\|_{\text{a}}$ in the direction of $\mathbf{b}$, i.e. $\tilde{\mathbf{x}} = \mathbf{x} \oplus \mathbf{b}/\|\mathbf{b}\|_{\text{a}}$, then the predictor becomes

$$\tilde{\eta}^{\text{comp}} = \langle \mathbf{b}, \tilde{\mathbf{x}} \rangle_{\text{a}} = \left\langle \mathbf{b}, \mathbf{x} \oplus \frac{\mathbf{b}}{\|\mathbf{b}\|_{\text{a}}} \right\rangle_{\text{a}} = \langle \mathbf{b}, \mathbf{x} \rangle_{\text{a}} + \frac{1}{\|\mathbf{b}\|_{\text{a}}} \langle \mathbf{b}, \mathbf{b} \rangle_{\text{a}} = \eta^{\text{comp}} + \|\mathbf{b}\|_{\text{a}}.$$

When $D = 3$, the estimated regression model can be visualized as a surface on a ternary diagram (van den Boogaart and Tolosana-Delgado, 2013). For $D > 3$, a graphical representation is not straightforward.

Further, we propose to perturb the composition in the direction of each component. This offers a new interpretation for the effect of altering the composition on the predictor. For example, a relative ratio change of $\alpha > 1$ (increase) or $\alpha < 1$ (decrease) in the first component of $\mathbf{x}$ with constant ratios of the remaining components can be achieved by perturbing the composition $\mathbf{x}$ by $\mathcal{C}(\alpha, 1, \ldots, 1)^{\mathrm{T}}$, i.e. $\tilde{\mathbf{x}} = \mathbf{x} \oplus \mathcal{C}(\alpha, 1, \ldots, 1)^{\mathrm{T}} = \mathcal{C}(\alpha x_1, x_2, \ldots, x_D)^{\mathrm{T}}$. This leads to a change in the predictor given by

$$\langle \mathbf{b}, \mathcal{C}(\alpha, 1, \ldots, 1)^{\mathrm{T}} \rangle_{\mathrm{a}} = \ln(\alpha) \left\{ \ln(b_1) - \frac{1}{D} \sum_{i=1}^{D} \ln(b_i) \right\} = \ln(\alpha) \, \mathrm{clr}_1(\mathbf{b}), \tag{8}$$

which is independent of the original composition $\mathbf{x}$. The effect of a relative increase in any of the components can hence best be understood by considering the clr-transform of $\mathbf{b}$, of which the elements sum to 0 and indicate the positive or negative effect of each component on the predictor. The difference in the predictor level between any two compositional data vectors $\mathbf{x}$ and $\mathbf{y}$ can be computed as

$$\langle \mathbf{b}, \mathbf{x} \ominus \mathbf{y} \rangle_{\mathrm{a}} = \sum_{i=1}^{D} \ln\left(\frac{x_i}{y_i} \right) \mathrm{clr}_i(\mathbf{b}),$$

which is a linear combination of the clr-co-ordinates of $\mathbf{b}$ with weights given by the component-wise log-ratios. A graphical representation of the effect of a compositional predictor can be made by visualizing $\mathrm{clr}(\mathbf{b})$ and comparing the elements with 0. Since $\boldsymbol{\beta} = \mathrm{ilr}(\mathbf{b}) = V^{\mathrm{T}} \ln(\mathbf{b}) = V^{\mathrm{T}} \mathrm{clr}(\mathbf{b})$ and $VV^{\mathrm{T}} = I_D - (1/D)\mathbf{1}_D \mathbf{1}_D^{\mathrm{T}}$, the clr-transform of $\mathbf{b}$ can be written as $\mathrm{clr}(\mathbf{b}) = V\boldsymbol{\beta}$. Confidence bounds can thus be constructed by using the corresponding covariance matrix $V\hat{\Sigma}V^{\mathrm{T}}$ where $\hat{\Sigma}$ is the estimated covariance matrix related to estimating $\boldsymbol{\beta}$. To quantify the influence of the compositional predictor on the level of the expected outcome in the Poisson and NB models, we exponentiate equation (8). The effect of a relative ratio change of α in component $i = 1, \ldots, D$ on the response scale is then given by $\alpha^{\mathrm{clr}_i(\mathbf{b})}$.

4.2.3. *Dealing with structural 0s in compositional predictors*

An additional difficulty when incorporating the compositional information as predictors in the analysis of the claim counts is the presence of proportions of a specific component that are exactly 0. In the division of the driven distance by road type, for instance, many insured drivers did not drive abroad during the policy period observed. Since compositional data are always analysed by considering log-ratios of the components (see Section 4.2.1), a workaround is necessary.

In the compositional data literature, different types of 0s are being distinguished (Pawlowsky-Glahn *et al.*, 2015). *Rounded 0s* occur when certain components may be unobserved because their true values are below the detection limit (cf. geochemical studies). *Count 0s* refer to zero values due to the limited size of the sample in compositional data arising from count data. In our setting, the zero values are truly 0 and are not due to imprecise or insufficient measurements. Such 0s are called *structural 0s*. The structural 0s patterns in the data set are listed in the appendix A of the on-line supplementary material. The presence of 0s is most prominent for splitting distance by road types as 40% of the drivers did not go abroad. 0s are most often dealt with by using replacement strategies (see for example Martín-Fernández *et al.* (2011), for an overview), which do not make sense for structural 0s. A general methodology is still to be developed (see for example Aitchison and Kay (2003) and Bacon Shone (2003)). In particular, a method does not exist that deals with compositional data with structural 0s as predictor in regression models. Applying the ilr-transform to the compositional data $\mathbf{x}$ and using the transformed $\mathbf{z}$ as explanatory variables in the predictor as discussed in Section 4.2.2 is no longer possible.

For an observation with structural 0s, we can only consider the *subcomposition* of non-zero components. We let $M \subset \{1, 2, \ldots, D\}$ denote the set of indices of the structural 0s of a composition $\mathbf{x}$. The subcomposition $\mathbf{x}_{M^c}$ of non-zero components $M^c = \{i_1, \ldots, i_m\} = \{1, 2, \ldots, D\} \setminus M$ of $\mathbf{x}$ is then obtained by applying the closure operation to the subvector $(x_{i_1}, \ldots, x_{i_m})$ of $\mathbf{x}$, i.e. $\mathbf{x}_{M^c} = \mathcal{C}(x_{i_1}, \ldots, x_{i_m}) = (x_{i_1}, \ldots, x_{i_m})/\Sigma_{i \in M^c} x_i$. The set of all possible structural 0 patterns M is denoted by $\mathcal{M}$. In the most general situation, $2^D - 1$ possible 0 patterns can occur when dealing with compositional data with D components (a structural 0 for every component being excluded). To indicate the 0 pattern of a composition $\mathbf{x}$ we introduce dummy variables

$$d_M(\mathbf{x}) = \begin{cases} 1 & \text{if the set of indices of the structural 0s of } \mathbf{x} \text{ is equal to } M, \\ 0 & \text{otherwise} \end{cases}$$

for all $M \in \mathcal{M}$. We propose two approaches to accommodate structural 0s in a regression context with compositional predictors, either via conditioning on the structural 0 pattern or via projection onto the orthogonal complement of the structural 0 parts.

4.2.3.1. Conditioning approach. We treat observations with different structural 0 patterns as qualitatively different subgroups within the data and model the effect of the compositional predictor conditional on the 0 pattern. The compositional predictor term η^{comp} of the regression model specifies a distinct effect $\mathbf{b}|_{M^c}$ for each 0 pattern:

$$\eta^{\text{comp}} = \sum_{M \in \mathcal{M}} d_M(\mathbf{x}) \langle \mathbf{b}|_{M^c}, \mathbf{x}_{M^c} \rangle_{\text{a}} = \sum_{M \in \mathcal{M}} d_M(\mathbf{x}) \langle \text{ilr}(\mathbf{b}|_{M^c}), \text{ilr}(\mathbf{x}_{M^c}) \rangle.$$

Conditionally on the 0 pattern M of the compositional data vector $\mathbf{x}$, the contribution to the predictor is given by the Aitchison inner product of the subcomposition $\mathbf{x}_{M^c}$ of non-zero components of $\mathbf{x}$ and a subcompositional simplicial gradient $\mathbf{b}|_{M^c}$ of the same dimension. The notation $\mathbf{b}|_{M^c}$ is used to indicate that the model parameters $\beta|_{M^c} = \text{ilr}(\mathbf{b}|_{M^c})$ differ by structural 0 pattern. Fitting this term requires us to compute for each compositional observation the ilr-transform $\text{ilr}(\mathbf{x}_{M^c})$ of the subcomposition of non-zero parts and to model the compositional predictor effect separately by 0 pattern. In the case of only one non-zero component, the Aitchison inner product is 0 and there is no contribution to the linear predictor. If deemed necessary a categorical effect based on the 0 pattern can be added to η^{comp}.

4.2.3.2. Projection approach. The compositional regression coefficients in the conditioning approach are different for each structural 0 pattern and hence only estimated by using observations with that particular 0 pattern. Instead of modelling the compositional predictor effect separately by 0 pattern, we alternatively propose a parsimonious simplification in which the regression parameters are shared across patterns.

For this, we regard a subcomposition $\mathbf{x}_{M^c}$ as an orthogonal projection of $\mathbf{x}$ that preserves the relative information that is contained in $\mathbf{x}_{M^c}$ and, simultaneously, filters out all the relative information involving parts in $\mathbf{x}_M$ (Pawlowsky-Glahn *et al.*, 2015). The clr-plane $\mathcal{H}^D$ is spanned by the non-orthogonal, non-basis vectors $\mathbf{w}_i = (-1/D, \ldots, -1/D, (D-1)/D, -1/D, \ldots, -1/D)$, for $i = 1, \ldots, D$, where the component equal to $(D-1)/D$ is placed at the ith component. The subcomposition $\mathbf{x}_{M^c}$ of the non-zero parts can be represented by an orthogonal projection P_M of the clr-transformed vector $\text{clr}(\mathbf{x})$ onto the null space of $\{\mathbf{w}_i, i \in M\}$, corresponding to the indices of the structural 0s of $\mathbf{x}$. van den Boogaart *et al.* (2006) showed that the projection P_M onto the orthogonal directions to M can be computed as

$$(P_M \operatorname{clr}(\mathbf{x}))_{M^c} = \operatorname{clr}(\mathbf{x}_{M^c})$$

and 0 otherwise. Hence, the subvector of P_M clr($\mathbf{x}$) that is related to the non-zero parts of $\mathbf{x}$ equals the clr-transform of the subcomposition $\mathbf{x}_{M^c}$ and the remaining elements of P_M clr($\mathbf{x}$) that are related to the structural 0s of $\mathbf{x}$ equal 0. We can express this projected clr-vector with respect to the chosen orthonormal ilr-basis and define $\mathbf{z} = V^T P_M$ clr($\mathbf{x}$) as a *generalized isometric log-ratio transformation* from the simplex (allowing for 0 components) to $\mathbb{R}^{D-1}$. In the case that $\mathbf{x}$ has no structural 0s, the generalized ilr-transform coincides with the regular ilr-transform (6). We suggest using these generalized ilr-co-ordinates in the compositional predictor and rewrite the term as

$$\begin{aligned}\eta^{\text{comp}} &= \sum_{i=1}^{D-1} \beta_i z_i = \operatorname{clr}(\mathbf{b})^T V V^T P_M \operatorname{clr}(\mathbf{x}) = \langle \operatorname{clr}(\mathbf{b}), P_M \operatorname{clr}(\mathbf{x}) \rangle = \langle (\operatorname{clr}(\mathbf{b}))_{M^c}, \operatorname{clr}(\mathbf{x}_{M^c}) \rangle \\ &= \langle \operatorname{clr}(\mathbf{b}_{M^c}), \operatorname{clr}(\mathbf{x}_{M^c}) \rangle = \langle \mathbf{b}_{M^c}, \mathbf{x}_{M^c} \rangle_a, \end{aligned} \tag{9}$$

where we used the fact that the elements of a clr-transform sum to 0 and that $(\operatorname{clr}(\mathbf{b}))_{M^c}$, the subvector of clr($\mathbf{b}$) that is related to the non-zero parts of $\mathbf{x}$, and $\operatorname{clr}(\mathbf{b}_{M^c})$, the clr-transform of the subcomposition of $\mathbf{b}$ related to the non-zero parts of $\mathbf{x}$, differ only by a vector of equal elements. Equation (9) shows that the compositional predictor in the projection approach is equivalent to the Aitchison inner product of the subcompositions of both $\mathbf{b}$ and $\mathbf{x}$ corresponding to the non-zero components of $\mathbf{x}$. In general, we can write

$$\eta^{\text{comp}} = \sum_{M \in \mathcal{M}} d_M(\mathbf{x}) \langle \mathbf{b}_{M^c}, \mathbf{x}_{M^c} \rangle_a.$$

Compared with the conditioning approach where the compositional regression coefficients $\mathbf{b}|_{M^c}$ are conditional on the 0 pattern, the projection approach is more parsimonious. The effect of each subgroup defined by the structural 0 patterns is obtained from the same model parameters $\boldsymbol{\beta}$, using subcompositions $\mathbf{b}_{M^c}$ of the corresponding compositional coefficient vector $\mathbf{b} = \operatorname{ilr}^{-1}(\boldsymbol{\beta})$. This simplifying assumption entails that leaving out the 0 components does not change the relative riskiness of the remaining components. Given an observation with structural 0s, the interpretation of the effect of a change to a non-zero component remains similar to before: if the relative ratio of the ith component of $\mathbf{x}_{M^c}$ changes by α, then the predictor changes by $\ln(\alpha)\ \operatorname{clr}_i(\mathbf{b}_{M^c})$. The clr-transformed subcomposition $\operatorname{clr}(\mathbf{b}_{M^c})$ can be obtained by recentring the parts of clr($\mathbf{b}$) corresponding to M^c around 0, i.e.

$$\operatorname{clr}(\mathbf{b}_{M^c}) = (\operatorname{clr}(\mathbf{b}))_{M^c} - \left\{ \frac{1}{m} \sum_{i \in M^c} \operatorname{clr}_i(\mathbf{b}) \right\} \mathbf{1}_m,$$

where $m = |M^c|$ is the number of non-zero parts in $\mathbf{x}$. Therefore, using the projection approach, a single graphical representation of clr($\mathbf{b}$) suffices to visualize and understand the effect of the compositional predictor term for each structural 0 pattern.

4.3. Model selection and assessment

Using the same form as Akaike's information criterion, AIC for a GAM is defined as

$$\text{AIC} = -2\hat{l} + 2\,\text{EDF} \tag{10}$$

where $\hat{l}$ is the log-likelihood, evaluated at the estimated model parameters obtained by using penalized likelihood maximization, and the effective degrees of freedom EDF are used instead of the actual number of model parameters. For details about the calculation of EDF see Wood *et al.* (2016). We used the implementation as available in the R package `mgcv` version 1.8-18.

As such, equation (10) measures the quality of the model as a trade-off between the goodness of fit and the model complexity.

For each of the four predictor sets (see Fig. 2(c)) variables are selected by AIC using an exhaustive search over all the possible combinations of variables given in Table 1. In our analysis, model selection is done without involving policyholder-specific random effects. All model specifications are estimated under both the Poisson and the NB framework. We restrict ourselves to additive regression models (i.e. no interactions) such that an exhaustive search is still feasible and the marginal effect of a single variable can be easily assessed, interpreted and visualized. Even though the 2011 EU ruling prohibits a distinction between men and women in car insurance pricing, we allow gender to be selected as a categorical predictor in the model. For the compositional predictors based on the different divisions of the distance driven, 10 structural 0 patterns occur for the road types, 20 for the time slots and three for week or weekend; see the appendix A of the on-line supplementary material. The model selection is performed separately for the conditioning and projection approach to the structural 0s. Following the projection approach, the three compositional predictor terms that we allow to be selected in the hybrid and telematics models are

$$\eta_{it}^{\text{comp}} = \sum_{M\in\mathcal{M}^{\text{road}}} d_M(\mathbf{x}_{it}^{\text{road}})\langle \mathbf{b}_{M^{\text{c}}}^{\text{road}}, \mathbf{x}_{it,M^{\text{c}}}^{\text{road}}\rangle_{\text{a}} + \sum_{M\in\mathcal{M}^{\text{time}}} d_M(\mathbf{x}_{it}^{\text{time}})\langle \mathbf{b}_{M^{\text{c}}}^{\text{time}}, \mathbf{x}_{it,M^{\text{c}}}^{\text{time}}\rangle_{\text{a}}$$
$$+ \sum_{M\in\mathcal{M}^{\text{week}}} d_M(\mathbf{x}_{it}^{\text{week}})\langle \mathbf{b}_{M^{\text{c}}}^{\text{week}}, \mathbf{x}_{it,M^{\text{c}}}^{\text{week}}\rangle_{\text{a}}.$$

Following the conditioning approach, the effect of the compositional predictor is modelled separately conditionally on the structural 0 pattern. However, on the basis of the relative frequencies of the 0 patterns in the data set, we allow only an additional compositional predictor term for the distinction by road type in the case that a car was not driven abroad, which occurs for 40% of the observations. All remaining 0 patterns are bundled into one residual group and their effect is modelled by using a categorical effect b_0; see Table A.4 of the on-line appendix A. Using the symbolic structural 0 pattern notation of the appendix A, the most comprehensive compositional predictor term in the conditioning approach can be denoted as

$$\eta_{it}^{\text{comp}} = d_{1111}(\mathbf{x}_{it}^{\text{road}})\langle \mathbf{b}|_{1111}^{\text{road}}, \mathbf{x}_{it,1111}^{\text{road}}\rangle_{\text{a}} + d_{1110}(\mathbf{x}_{it}^{\text{road}})\langle \mathbf{b}|_{1110}^{\text{road}}, \mathbf{x}_{it,1110}^{\text{road}}\rangle_{\text{a}} + \{1 - d_{1111}(\mathbf{x}_{it}^{\text{road}})$$
$$- d_{1110}(\mathbf{x}_{it}^{\text{road}})\} b|_0^{\text{road}} + d_{11111}(\mathbf{x}_{it}^{\text{time}})\langle \mathbf{b}|_{11111}^{\text{time}}, \mathbf{x}_{it,11111}^{\text{time}}\rangle_{\text{a}} + \{1 - d_{11111}(\mathbf{x}_{it}^{\text{time}})\} b|_0^{\text{time}}$$
$$+ d_{11}(\mathbf{x}_{it}^{\text{week}})\langle \mathbf{b}|_{11}^{\text{week}}, \mathbf{x}_{it,11}^{\text{week}}\rangle_{\text{a}} + \{1 - d_{11}(\mathbf{x}_{it}^{\text{week}})\} b|_0^{\text{week}}.$$

Predictive performance of these models is assessed by using *proper scoring rules* for count data; Table 2 (Czado *et al.*, 2009). Scoring rules assess the quality of probabilistic forecasts through a numerical score $s(P,n)$ based on the predictive distribution P and the observed count n. Lower scores indicate a better quality of forecast. A scoring rule is proper (Gneiting and Raftery, 2007) if $s(Q,Q) \leqslant s(P,Q)$ for all P and Q with $s(P,Q)$ the expected value of $s(P,\cdot)$ under Q. In general, we define by $p_k = \mathbb{P}(N=k)$ and $P_k = \mathbb{P}(N \leqslant k)$ the probability mass function and cumulative probability function of the predictive distribution P for count variable N. The probability mass at the observed count n is denoted as p_n. The mean and standard deviation of P are written as μ_P and σ_P respectively, and we set $\|p\| = \Sigma_{k=0}^{\infty} p_k^2$.

We compare the predictive performance of the best models according to AIC under the four predictor sets, with or without offset in the predictor (1), and using a Poisson or NB distribution. We apply the proper scoring rules to the predictive count distributions of the observed claim counts. We adopt a K-fold cross-validation approach (Hastie *et al.*, 2009) with $K=10$ and apply the same partition to assess each model specification. Let $\kappa_{it} \in \{1,2,\ldots,K\}$ be the part of the

Table 2. Proper scoring rules for count data

Score	*Formula*
Logarithmic	$\text{logs}(P,n) = -\log(p_n)$
Quadratic	$\text{qs}(P,n) = -2p_n + \|p\|$
Spherical	$\text{sphs}(P,n) = -p_n/\|p\|$
Ranked probability	$\text{rps}(P,n) = \Sigma_{k=0}^{\infty}\{P_k - \mathbf{1}(n \leqslant k)\}^2$
Dawid–Sebastiani	$\text{dss}(P,n) = \left(\frac{n-\mu_P}{\sigma_P}\right)^2 + 2\log(\sigma_P)$
Squared error	$\text{ses}(P,n) = (n-\mu_P)^2$

data to which the observed claim count n_{it} of policyholder i in policy period t is allocated by the randomization. Denote by $\hat{P}_{it}^{-\kappa_{it}}$ the predictive count distribution for observation n_{it} estimated without the κ_{it}th part of the data. The K-fold cross-validation score $\text{CV}(s)$ is then given by

$$\text{CV}(s) = \frac{1}{\sum_{i=1}^{I} T_i} \sum_{i=1}^{I} \sum_{t=1}^{T_i} s(\hat{P}_{it}^{-\kappa_{it}}, n_{it}),$$

where s is any of the aforementioned proper scoring rules and smaller values of $\text{CV}(s)$ indicate better forecasts.

5. Results

All computations are performed with R version 3.4.1 (R Core Team, 2017) and, in particular, R package `mgcv` version 1.8-18 (Wood, 2011) is used for the parameter estimation in the GAMs and `compositions` version 1.40-1 (van den Boogaart *et al.*, 2014) to compute the transformations of the compositional data. For brevity and clarity of this presentation, we show only the results (tables and figures) for the Poisson models using the projection approach for the structural 0s and highlight differences from the NB models and the conditioning approach, if any. As supplementary material, we provide accompanying R code which illustrates how to apply the methods that are presented in this work on simulated data with a similar structure. Following either the conditioning or the projection approach to handle structural 0s, we demonstrate how to include a compositional predictor in a GAM and how to visualize the effect.

5.1. Model selection

The variables that were selected for each of the predictor sets were identical for the Poisson and NB models; Table 3. The functional forms of the best models selected are given in the appendix B of the on-line supplementary material. The offset versions of the classical and time hybrid model replace the term $f_1(\text{time}_{it})$ by $\ln(\text{time}_{it})$, without any regression coefficient in front. This causes the expected number of reported MTPL claims, $\mu_{it} = \mathbb{E}(N_{it}) = \exp(\eta_{it})$, to be proportional to the duration of the policy period. In the offset versions of the metre hybrid and telematics model, the flexible term related to distance is replaced by an offset $\ln(\text{distance}_{it})$, imposing the risk to be proportional to the distance.

The models which are allowed to use the policyholder information prefer the use of experience, measured as the years since obtaining the driver's licence, instead of age to segment the risk in

Table 3. Variables contained in the best Poisson model for each of the predictor sets by using the projection approach for structural 0s†

Predictor	*Classical*		*Time hybrid*		*Metre hybrid*		*Telematics*	
Policy								
time	×	Offset	×	Offset				
age								
experience	×	×	×	×	×	×		
gender	×	×						
material	×	×	×	×	×	×		
postal code	×	×	×	×	×	×		
bonus-malus	×	×	×	×	×	×		
age vehicle	×	×	×	×	×	×		
kwatt			×	×	×	×		
fuel	×	×	×		×			
Telematics								
distance					×	Offset	×	Offset
yearly distance			×	×				
average distance			×	×	×	×	×	×
road type			×	×	×	×	×	×
time slot			×	×	×	×	×	×
week/weekend								×

†The second column of each predictor set refers to the model with an offset for either time or metre. The best NB models were identical to the best Poisson models.

young drivers. Gender is selected as an important covariate in only the classical models, not in any of the hybrid models, indicating that the telematics information renders the use of gender as a rating variable redundant. In the offset variants of both hybrid models the fuel-term is dropped. The newly introduced telematics predictors road type and time slot are selected in both the hybrid and the telematics models. The week/weekend-term is selected in only the offset variant of the telematics model.

The second-best models, with only a slightly higher AIC-value, show that adding kwatt to the classical model gives a comparable model fit and the same holds for adding week/weekend to the hybrid and telematics models. Furthermore, fuel can easily be left out of the hybrid models without deteriorating the fit.

Using the conditioning approach for structural 0s as opposed to the projection approach, the main difference in the selection of variables is that the compositional predictor week/weekend is always included in the hybrid and telematics models. Both the 1111 and the 1110 0 patterns of road type are selected. Stepwise adding more 0 patterns of road type to the predictor did not improve AIC, and similarly for the division by time slot and week/weekend. The variables selected were again identical under the Poisson and NB model specification.

For each of these best model formulations, we added a policyholder-specific random effect in predictor (1) to account for possible dependence from observing policyholders over multiple policy periods. However, none of the added random effects were deemed necessary at the 5% level of significance by using the approximate test of Wood (2013).

5.2. Model assessment

Table 4 reports AIC and all six proper scoring rules obtained by using tenfold cross-validation for each predictor set under the Poisson model specification using the projection approach for structural 0s. These performance tools unanimously indicate that the time hybrid model without

Table 4. Model assessment of the best models according to AIC for each of the four predictor sets under the Poisson model specification and using the projection approach for structural 0s†

Predictor set	*Offset*	*EDF*	*AIC*		*logs × 10*		*qs × 10*		*sphs × 10*		*rps × 10^2*		*dss*		*ses × 10^2*	
			Value	*Rank*	*Value*	*Rank*	*Value*	*Rank*	*Value*	*Rank*	*Value*	*Rank*	*Value*	*Rank*	*Value*	*Rank*
Classical	No	32.15	11896	6	1.790	6	−9.1858	6	−9.5822	6	4.224	6	−2.206	5	4.535	6
	Yes	27.27	11995	8	1.804	8	−9.1838	8	−9.5816	8	4.234	8	−2.129	6	4.547	8
Time hybrid	No	34.11	11734	1	1.766	1	−9.1909	1	−9.5837	1	4.196	1	−2.266	1	4.502	1
	Yes	30.41	11811	3	1.777	3	−9.1891	3	−9.5831	3	4.205	3	−2.212	4	4.512	3
Metre hybrid	No	34.32	11743	2	1.767	2	−9.1907	2	−9.5836	2	4.197	2	−2.259	2	4.503	2
	Yes	30.37	11866	5	1.785	5	−9.1884	4	−9.5829	4	4.209	4	−2.007	7	4.517	4
Telematics	No	15.05	11862	4	1.784	4	−9.1871	5	−9.5826	5	4.216	5	−2.226	3	4.526	5
	Yes	11.43	11989	7	1.803	7	−9.1850	7	−9.5820	7	4.228	7	−1.965	8	4.538	7

†The second row of each predictor set refers to the model with an offset for either time or metre. For each model we list the effective degrees of freedom EDF, Akaike information criterion AIC and six cross-validated proper scoring rules: logarithmic, logs, quadratic, qs, spherical, sphs, ranked probability, rps, Dawid–Sebastiani, dss, and squared error scores, ses. For AIC and the proper scoring rules, the first column represents the value and the second column the rank.

offset scores best. The metre hybrid model is a close second. Their respective versions with an offset and the telematics model without offset conclude the top five according to all criteria except for the Dawid–Sebastiani score. This demonstrates the significant influence of telematics-constructed variables on the predictive power of the model. In addition, the telematics model without offset outperforms the classical models across all assessment criteria. Hence, using only telematics predictors is considered to be better than the use of the traditional rating variables.

Across all predictor sets, the use of an offset for the exposure to risk, either time or metre, is too restrictive for these data. From a statistical point of view, the time or metre rating unit cannot be considered to be directly proportional to the risk. However, from a business point of view, it is convenient to consider a proportional approach because of its simplicity and explainability.

Similar results are obtained under the NB model specification and using the conditioning approach for structural 0s. The rankings of the predictor sets according to AIC are the same as in Table 4 under the NB model specification and/or using the conditioning approach. The AIC-values for each predictor set under the NB model specification compared with their Poisson counterpart were slightly higher for the classical and hybrid models and slightly lower for the telematics models, indicating that only the telematics predictor sets benefit from the additional parameter to capture overdispersion. The model assessment using proper scoring rules led to the same conclusions as before.

Beside an exhaustive search among additive terms, we have explored the use of interactions between categorical, between continuous, between categorical and continuous, and between categorical and compositional predictors. Slight marginal improvements in AIC could only be achieved in the classical model by further refining the effects of experience, age vehicle and material by gender without changing the rankings of the best models in Table 4.

5.3. Visualization and discussion

The effects of each predictor variable in the best time hybrid model without an offset are graphically displayed in Fig. 5 for the policy variables and in Fig. 6 for the telematics variables. By exponentially transforming the additive effects, we show the multiplicative effects on the ex-

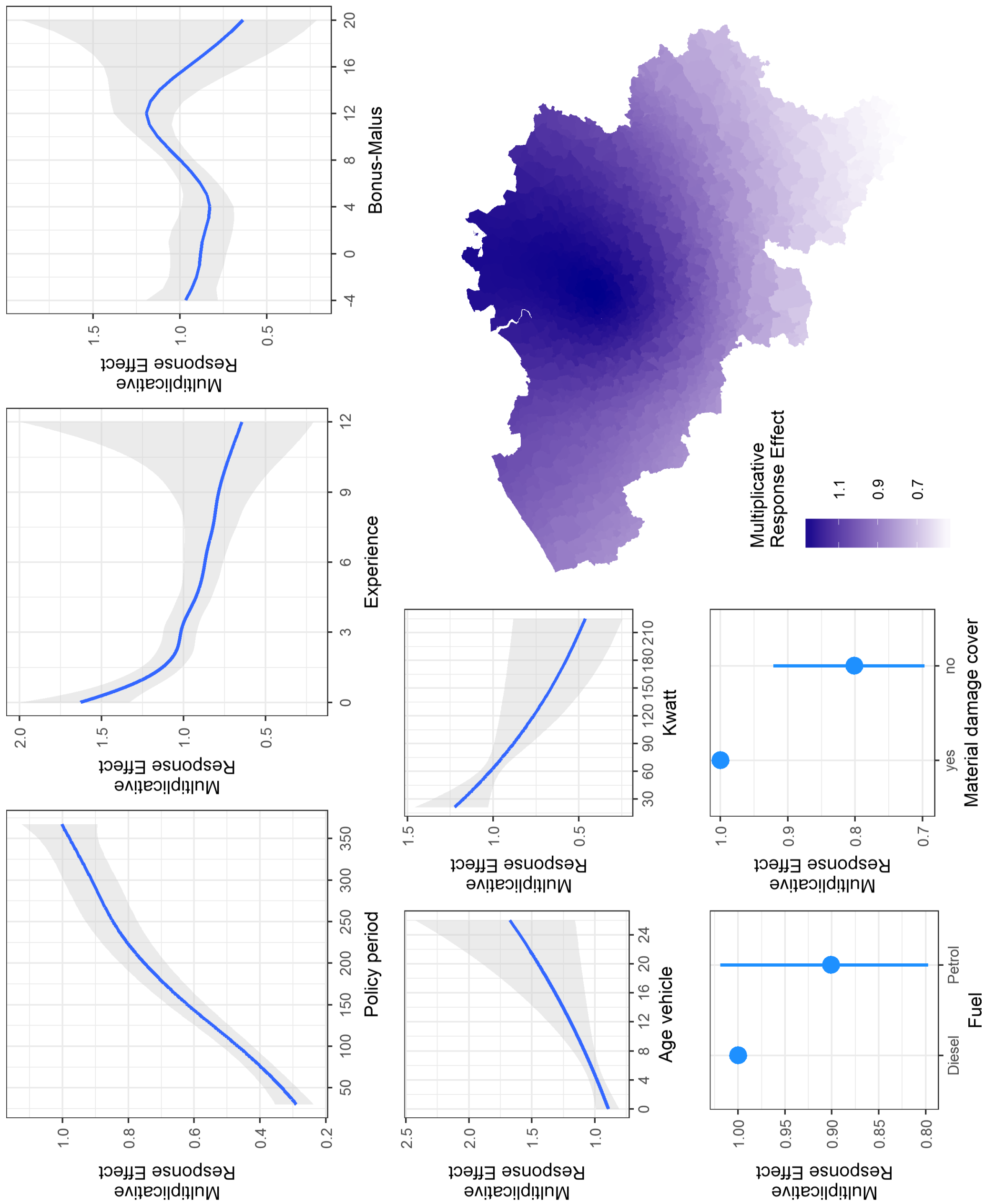


Fig. 5. Multiplicative response effects of the policy predictor variables of the time hybrid model

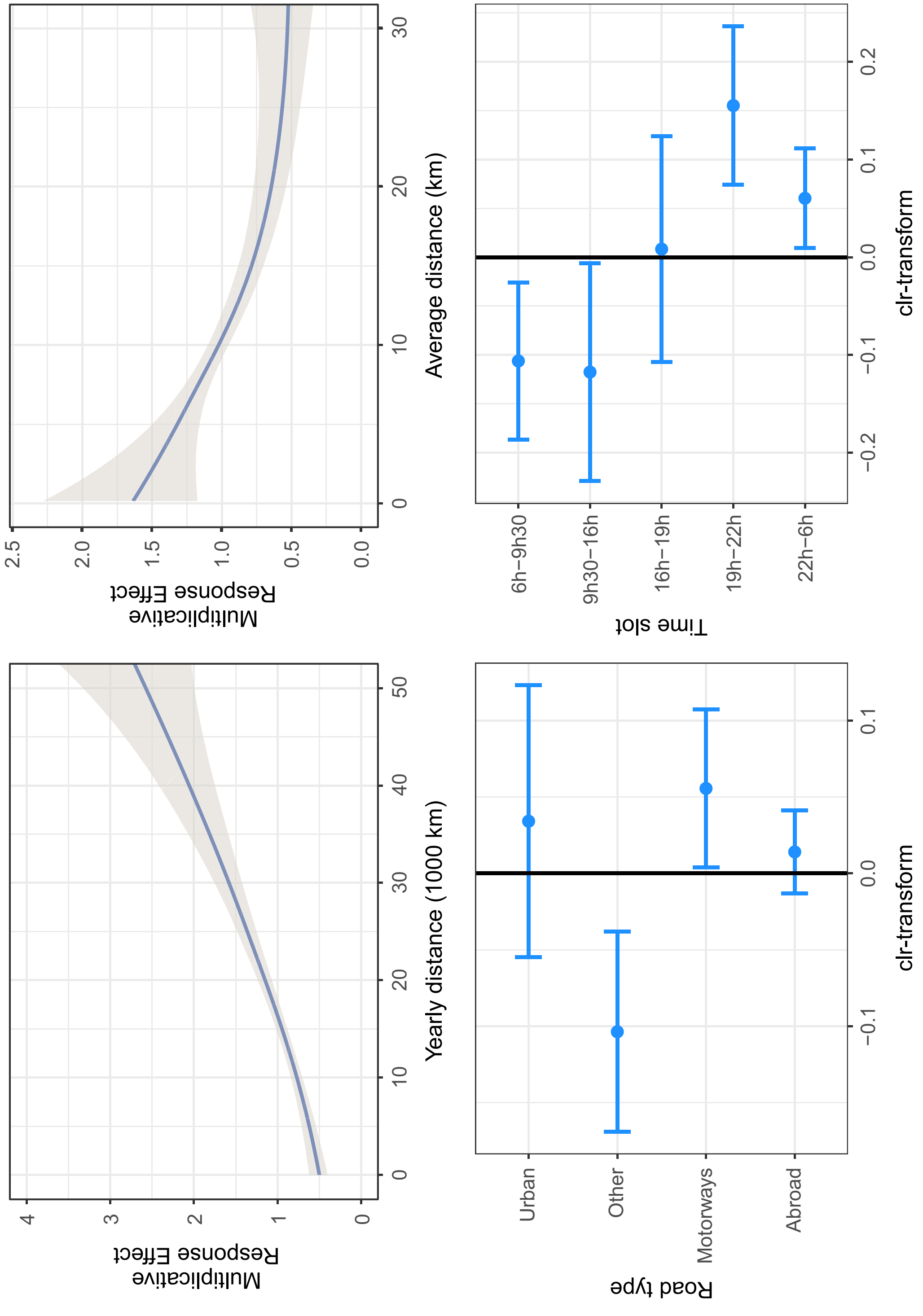


Fig. 6. Multiplicative response effects of the telematics predictor variables of the time hybrid model

pected number of claims for each categorical parametric, continuous smooth or geographical term in the fitted model. For the categorical predictors we quantify the uncertainty of those estimates by constructing individual 95% confidence intervals based on the large sample normality of the model parameter estimators. Bayesian 95% confidence pointwise intervals are used for the smooth components of the GAM and include the uncertainty about the intercept (Marra and Wood, 2012). For the compositional data predictors, we visualize the clr-transform of the corresponding model parameters with 95% confidence intervals along with a reference line at zero (see Section 4.2.2). In the on-line supplementary material, similar graphs for the other three predictor sets (see Fig. 2(c)) are shown in the appendix C and the relative importance of these predictors is quantified and visualized in the appendix D of the on-line supplementary material. In the remainder of this section, we discuss the insights and the interpretations for both the policy and the telematics variables in each of these models.

5.3.1. Policy variables

The rating unit policy period in the classical and time hybrid models always has a monotone increasing estimated effect. The longer a policyholder is insured, the higher the premium amount, *ceteris paribus*. Using the fact that the level of the non-linear smooth component is not uniquely identifiable (see Section 4.1), we vertically translated the estimated smooth term to pass the point (365, 0) on the predictor scale (and hence (365, 1) on the response scale) for ease of interpretation.

The smooth effect of experience embodies the higher risk that is posed by younger, less experienced drivers. The increased risk is more outstanding in the first two years for the hybrid models compared with the classical model.

In the classical model, the significant effect of gender indicates that women are 16% less risky drivers than men. However, when telematics predictors are taken into account in the hybrid models, the categorical variable gender is no longer selected as a predictor. Nor did any interaction term between gender and a categorical, a continuous or a compositional predictor improve AIC. The perceived difference between women and men can hence be explained through differences in driving habits. In particular, female drivers in the portfolio drive significantly fewer kilometres yearly compared with men (15409 *versus* 18570 km on average, with a p-value smaller than 0.001 by using a two-sample t-test). Similar findings were reported in Ayuso *et al.* (2016a, b). In light of the EU rules on gender neutral pricing in insurance, this shows how moving towards car insurance rating based on individual driving habits and style can resolve possible discrimination of basing the premium on proxies such as gender.

The smooth effects of BM in the classical and hybrid models are non-linear and somewhat counterintuitive. Given the lack of a lengthy claim history of the young drivers of this portfolio, the BM levels of the insured drivers are not yet fully developed and stabilized. The majority of the drivers have a BM level between 4 and 12 for which the effect on claim frequency is increasing. For the highest BM levels, however, the effect is declining, albeit with a high uncertainty due to a lack of observations in this region. Furthermore, the effect does not decrease for the lowest BM levels. This can be explained by an improper use of the BM scale as a marketing tool to attract new customers. By lowering the initial value of the BM scale, the insurer can reduce the premium that a potential new policyholder must pay.

When it comes to characteristics of the car, insured drivers driving older vehicles have an estimated higher risk of accidents. The smooth effect of age vehicle is estimated as a straight line on the predictor scale in the classical and hybrid models. The effect of kwatt in the hybrid models also reduced to a straight line on the predictor scale. When the insured vehicle has more horsepower, the estimated expected number of claims is lower. The categorical predictor fuel shows that vehicles using petrol have an estimated lower risk of accidents compared with diesel vehicles. This difference is, however, smaller and no longer statistically significant in the hybrid models compared with the classical model, where it serves as a proxy variable for the distance driven. Indeed, vehicles using diesel as opposed to petrol are driven significantly more kilometres in our portfolio (18 940 *versus* 13 267 yearly on average).

In both the classical and the hybrid models, the policies without material damage cover have a 20% lower estimated expected number of claims. This may be explained by the reluctance of some insured drivers without additional material damage coverage to report small accidents. Because BM mechanisms are independent of the claim amount, filing a claim leads to premium surcharges which may be more disadvantageous for policyholders than for them to defray the third-party costs. This phenomenon is known as the hunger for bonus (Denuit *et al.*, 2007). Insured drivers with an additional material damage cover are less inclined to do so since their own, first-party costs are also covered, making it more worthwhile to report a claim at fault. Including telematics variables in the model does not affect this discrepancy.

The geographical effect (postal code), plotted on top of a map of Belgium for the classical and hybrid models, captures the remaining spatial heterogeneity based on the postal code where the policyholder resides. For the classical model, the graph shows higher claim frequencies for urban areas like Brussels in the middle, Antwerp in the north and Liège in the east and lower claim frequencies in the more sparsely populated regions in the south. The geographic variation, however, decreases strongly in the hybrid models because of the inclusion of telematics predictors that are not taken into account in the classical model. The EDF corresponding to the spatial smooth reduced from 15.8 in the classical model to 4.1 and 4.4 in the time and metre hybrid model respectively. This is satisfactory as it means that, instead of overrelying on geographical proxies, the hybrid models are basing the insurance premium on actual differences in driving habits, which are more closely related to the accident risk.

5.3.2. Telematics variables

In the metre hybrid and telematics models, distance is used as the rating unit. Similar to the time effect in the classical and time hybrid model, the effect of the risk exposure is estimated as a monotone increasing function. The accident risk, however, does not vanish for insured drivers who drive hardly any kilometres during the observation period.

The yearly distance is used in the time hybrid model, which uses time as exposure, to differentiate between drivers who travel many *versus* few kilometres yearly. In this way, the distance driven is rescaled yearly (see Section 3.2) and used as an additional risk factor having a weaker effect on the claim frequency compared with the metre hybrid and telematics models where distance is used as a rating unit. In both hybrid models and the telematics model, the estimated average distance effect shows lower claim frequencies for insured drivers who on average drive long distances.

Our modelling approach using compositional predictors separates on the one hand the effect of an overall increase in the driven distance and on the other hand the effect of a change in the division of the driven distance into different categories. This allows us qualitatively and quantitatively to interpret and visualize the effect of individual driving habits on the expected claim frequencies.

The clr-transforms of the model coefficients related to the compositional road type predictor in the telematics model show how insured drivers who drive relatively more on urban roads have higher claim frequencies and insured drivers who drive relatively more on the ‘other’ road type have lower claim frequencies. In the hybrid models, these effects head in the same direction with the exception that motorways are perceived as riskier. The elevated accident risk

$$\begin{tikzcd}
\mathbb{R}^3 & \mathcal{S}^4 & \mathcal{H}^4 \subset \mathbb{R}^4 \\
\boldsymbol{\beta} = (-0.097, 0.074, 0.016)^{\mathrm{T}} & \boldsymbol{b} = (0.258, 0.225, 0.264, 0.253)^{\mathrm{T}} \arrow[l, "\mathrm{ilr}"'] \arrow[r, "\mathrm{clr}"] & \mathrm{clr}(\boldsymbol{b}) = (0.034, -0.104, 0.056, 0.014)^{\mathrm{T}} \\
\boldsymbol{z} = (-0.318, -0.275, -2.231)^{\mathrm{T}} \arrow[r, "\mathrm{ilr}^{-1}"'] & \boldsymbol{x} = (0.442, 0.282, 0.252, 0.024)^{\mathrm{T}} & \mathrm{clr}(\boldsymbol{x}) = (0.981, 0.532, 0.419, -1.932)^{\mathrm{T}} \arrow[l, "\mathrm{clr}^{-1}"]
\end{tikzcd}$$

Fig. 7. Representations in the ilr-, simplex and clr-space of the estimated regression parameters in the time hybrid model with respect to the compositional predictor term road type (with components *urban*, *other*, *motorways* and *abroad*) and the average compositional data vector without structural 0s

for insured drivers driving more on urban roads is in line with Paefgen *et al.* (2014), where the driven distance is divided over 'highway', 'urban' and 'extra-urban' road types. They, however, neglected the compositional nature of this predictor in the analysis and did not incorporate any of the classical policy risk factors in the logistic regression model. In Ayuso *et al.* (2014), the percentage of urban driving was considered an important variable to predict either the time or the distance to the first accident, although percentages driven on different types of road were not considered. Using either a quadratic effect or a categorical effect (urban driving $>25\%$) in Weibull regression models shows how increased percentages of urban driving reduce both the expected time and the distance to the first accident.

The estimated model coefficients $\boldsymbol{\beta}^{\mathrm{road}}$ of the compositional road type predictor and the corresponding simplicial gradient $\mathbf{b}^{\mathrm{road}}$ and clr-transform $\mathrm{clr}(\mathbf{b}^{\mathrm{road}})$ are presented in Fig. 7. On the basis of the latter we can quantitatively interpret the effect of road type in the time hybrid model. For instance, a relative ratio increase of 50% to the 'other' road type component, with constant ratios of the remaining components, results in a multiplicative decrease of $1.50^{-0.104} = 0.959$ for the expected number of claims. Applied to the compositional road type data vector $\mathbf{x} = (0.442, 0.282, 0.252, 0.024)$ of Fig. 7, this relative ratio increase to the 'other' road type component would change the compositional vector to $\tilde{\mathbf{x}} = \mathbf{x} \oplus \mathcal{C}(1, 1.50, 1, 1)^{\mathrm{T}} = (0.387, 0.371, 0.221, 0.021)^{\mathrm{T}}$. On the basis of our projection approach for structural 0s, the interpretation is similar for a relative ratio change to a non-zero component of observations with a certain 0 pattern. In particular, for someone who did not drive abroad we base the interpretation of the effect on the clr-transform of the related subcomposition, i.e. $\mathrm{clr}(\mathbf{b}^{\mathrm{road}}_{1110}) = (0.038, -0.099, 0.061)^{\mathrm{T}}$, obtained by recentring $\mathrm{clr}(\mathbf{b}^{\mathrm{road}})$ without the abroad component near zero.

The compositional time slot predictor in the hybrid and telematics models indicates that policyholders who drive relatively more in the morning have lower claim frequencies and policyholders who drive relatively more in the evening and during the night have higher claim frequencies. For instance, the multiplicative response effect of a relative ratio increase of 50% in the evening component (between 7 p.m. and 10 p.m.) is equal to $1.50^{0.155} = 1.065$. In Paefgen *et al.* (2014), the accident risk is considered to be lower during the daytime (between 5 a.m. and 6 p.m.) compared with the evening (between 6 p.m. and 9 p.m.), based on the estimated coefficients of linear model terms of the log-transformed percentages of the distance driven in these time slots. Ayuso *et al.* (2014) reported how a higher percentage of driving at night reduces the expected time to a first accident, where the effect was modelled linearly, with no further distinction in time slots.

6. Conclusion

Telematics insurance offers new opportunities for insurers to differentiate drivers on the basis of their driving habits and style. By aggregating the telematics data at the level of the policy period by policyholder and combining them with traditional policy(holder) rating variables, we construct predictive models for the frequency of MTPL claims at fault. GAMs with a Poisson or NB response are used to model the effects of predictors in a smooth, yet interpretive way. The divisions of the driven distance into four road types and five time slots forms a challenge from a methodological point of view that has not been addressed in the literature. We demonstrate how to include this information as compositional predictors in the regression and formulate a new way of how to interpret their effect on the average claim frequency.

Our research reveals the significant influence of the use of telematics data through an exhaustive model selection and an assessment of the predictive performance. The time hybrid is the best model according to AIC and all proper scoring rules, closely followed by the metre hybrid model. The model using only telematics variables is ranked higher than the best classical model using only traditional policy information.

The compositional predictors show that a further classification of the distance driven based on the location and the time is relevant. Our contribution indicates that driving more on urban roads or motorways and in the evening or at night contributes to a riskier driving pattern. The best hybrid models highlight that certain popular pricing factors (gender, fuel and postal code) are indeed proxies for the driving habits and part of their predictive power is taken over by the distance driven and the splits into different categories. Hence, we demonstrate by using careful statistical modelling how the use of telematics variables is an answer to the EU regulation on insurance pricing practices that bans the use of gender as a rating factor.

In the case of multiple insured drivers, it is unclear which characteristics (such as age, experience and gender) the insurer must use to determine the premium. We proceed, in consultation with the Belgian insurer who provided the data, by identifying the driver with the lowest experience as the main driver and use his policyholder information as predictors in the regression for tarification purposes. In practice, when a parent adds a child as a driver in the policy, a premium surcharge is often waived to prevent the policyholder from lapsing. By shifting towards pricing based on telematics information as we do in this research, this tarification issue becomes less of a problem because the premium will be usage based.

Pricing using telematics data can be seen as falling in between *a priori* and *a posteriori* pricing. The driving habits and style are not traditional *a priori* variables since they cannot be determined before the policyholder starts to drive. Insurers now reason that available UBI products are purchased only by drivers who consider themselves to be either safe or low kilometre drivers. This potential form of positive selection, which could not be quantified on the basis of the studied portfolio alone, validates an upfront discount on the traditional insurance premium. On the basis of the telematics data collected over time, insurers can set up a discount structure to adapt the premium in an *a posteriori* way. The discount structure can depend on the actual distance driven, with a further personalized differentiation based on the riskiness of the profile as perceived from the driving habits of the insured person. The insights that are provided in this paper reveal which elements can be adopted in such a structure, for instance, by making kilometres driven on urban roads, or in the evening or at night more expensive.

In conclusion, telematics technology provides means to insurers to align premiums with risk better. Pay as you drive insurance is a first step in which the number of driven kilometres, the type of road and the time of day are combined with the traditional self-reported information such as policyholder and car characteristics to calculate insurance premiums. A next step is pay how you drive insurance, where on top of these driving habits also the driving style is considered to assess how risky someone drives by monitoring for instance speed limit infringements, harsh braking, excessive acceleration and cornering style. The ideas and statistical framework that were presented can be extended to incorporate such additional pay how you drive predictors if they are available.

Acknowledgements

The authors are grateful to the Associate Editor and their referees for their valuable comments and suggestions. Support is acknowledged from the agency for Innovation by Science and Technology (grant IWT 131173) and the Flemish Supercomputer Centre (Vlaams Supercomputer Centrum). Furthermore, Katrien Antonio acknowledges financial support from the Ageas Continental Europe Research Chair at KU Leuven and from KU Leuven's research council (project COMPACT C24/15/001). Gerda Claeskens acknowledges support through the Interuniversity Attraction Pole Research Network P6/03 of the Belgian State (Belgian research policy) and a KU Leuven grant (GOA/12/14). The authors also thank their contact person at the insurance company for the smooth co-operation.

References

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Aitchison, J. and Kay, J. W. (2003) Possible solution of some essential zero problems in compositional data analysis. In *Proc. 1st Compositional Data Analysis Wrkshp, Girona* (eds S. Thió-Henestrosa and J. A. Martín-Fernández).

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

Ayuso, M., Guillén, M. and Pérez-Marín, A. M. (2014) Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accid. Anal. Prevn*, **73**, 125–131.

Ayuso, M., Guillén, M. and Pérez-Marín, A. M. (2016a) Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, **4**, no. 2.

Ayuso, M., Guillén, M. and Pérez-Marín, A. M. (2016b) Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transprtn Res.* C, **68**, 160–167.

Bacon Shone, J. (2003) Modelling structural zeros in compositional data. In *Proc. 1st Compositional Data Analysis Wrkshp, Girona* (eds S. Thió-Henestrosa and J. A. Martín-Fernández).

Barceló-Vidal, C., Martín-Fernández, J. A. and Mateu-Figueras, G. (2011) Compositional differential calculus on the simplex. In *Compositional Data Analysis: Theory and Applications* (eds V. Pawlowsky-Glahn and A. Buccianti). Chichester: Wiley.

van den Boogaart, K. G. and Tolosana-Delgado, R. (2013) *Analyzing Compositional Data with R*. Berlin: Springer.

van den Boogaart, K., Tolosana-Delgado, R. and Bren, M. (2006) Concepts for handling zeroes and missing values in compositional data. In *Proc. 11th A. Conf. International Association for Mathematical Geology, Liège* (ed. E. Pirard).

van den Boogaart, K. G., Tolosana, R. and Bren, M. (2014) compositions: compositional data analysis. *R Package Version 1.40-1*. Technische Universität Bergakademie Freiberg, Freiberg. (Available from `https://CRAN.R-project.org/package=compositions`.)

Bordoff, J. E. and Noel, P. J. (2008) Pay-as-you-drive auto insurance: a simple way to reduce driving-related harms and increase equity. *Discussion Paper*. Brookings Institution, Washington DC.

Boucher, J.-P., Pérez-Marín, A. M. and Santolino, M. (2013) Pay-as-you-drive insurance: the effect of the kilometers on the risk of accident. *An. Inst. Act. Espan.*, **19**, 135–154.

Butler, P. (1993) Cost-based pricing of individual automobile risk transfer: car-mile exposure unit analysis. *J. Act. Pract.*, **1**, 51–84.

Czado, C., Gneiting, T. and Held, L. (2009) Predictive model assessment for count data. *Biometrics*, **65**, 1254–1261.

Denuit, M. and Lang, S. (2004) Non-life ratemaking with Bayesian GAMs. *Insur. Math. Econ.*, **35**, 627–647.

Denuit, M., Marechal, X., Pitrebois, S. and Walhin, J. (2007) *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-malus Systems*. Chichester: Wiley.

Desyllas, P. and Sako, M. (2013) Profiting from business model innovation: evidence from pay-as-you-drive auto insurance. *Res. Poly.*, **42**, 101–116.

Economist (2013) How's my driving? *Economist*, Feb. 23rd. (Available from `http://econ.st/Yd5x3C`.)

Egozcue, J. J., Barceló-Vidal, C., Martín-Fernández, J. A., Jarauta-Bragulat, E., Díaz-Barrero, J. L. and Mateu-Figueras, G. (2011) Elements of simplicial linear algebra and geometry. In *Compositional Data Analysis: Theory and Applications* (eds V. Pawlowsky-Glahn and A. Buccianti). Chichester: Wiley.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barcelo-Vidal, C. (2003) Isometric logratio transformations for compositional data analysis. *Math. Geol.*, **35**, 279–300.

Ferreira, J. and Minikel, E. (2010) Pay-as-you-drive auto insurance in Massachusetts: a risk assessment and report on consumer. Massachusetts Institute of Technology, Cambridge. (Available from `http://mit.edu/jf/www/payd/PAYD_CLF_Study_Nov2010.pdf`.)

Filipova-Neumann, L. and Welzel, P. (2010) Reducing asymmetric information in insurance markets: cars with black boxes. *Telem. Inform.*, **27**, 394–403.

Frees, E. W. (2014) Frequency and severity models. In *Predictive Modeling Applications in Actuarial Science*, vol. 1 (eds E. W. Frees, R. A. Derrig and G. Meyers). Cambridge: Cambridge University Press.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction and estimation. *J. Am. Statist. Ass.*, **102**, 359–378.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. London: Chapman and Hall.

Greenberg, A. (2009) Designing pay-per-mile auto insurance regulatory incentives. *Transprtn Res.* D, **14**, 437–445.

Hastie, T. and Tibshirani, R. (1986) Generalized additive models. *Statist. Sci.*, **1**, 297–318.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edn. New York: Springer.

Hron, K., Filzmoser, P. and Thompson, K. (2012) Linear regression with compositional explanatory variables. *J. Appl. Statist.*, **39**, 1115–1128.

de Jong, P. and Heller, G. (2008) *Generalized Linear Models for Insurance Data*. Cambridge: Cambridge University Press.

Klein, N., Denuit, M., Lang, S. and Kneib, T. (2014) Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale and shape. *Insur. Math. Econ.*, **55**, 225–249.

Krivobokova, T. and Kauermann, G. (2007) A note on penalized spline smoothing with correlated errors. *J. Am. Statist. Ass.*, **102**, 1328–1337.

Lancaster, P. and Salkauskas, K. (1986) *Curve and Surface Fitting: an Introduction*. London: Academic Press.

Lemaire, J., Park, S. C. and Wang, K. C. (2016) The use of annual mileage as a rating variable. *Astin Bull.*, **46**, 39–69.

Litman, T. (2011) Distance-based vehicle insurance feasibility, costs and benefits. Victoria Transport Policy Institute, Victoria. (Available from `http://www.vtpi.org/dbvi_com.pdf`.)

Litman, T. (2015) Pay-as-you-drive vehicle insurance: converting vehicle insurance premiums into use-based charges. Victoria Transport Policy Institute, Victoria. (Available from `http://www.vtpi.org/tdm/tdm79.htm`.)

Marra, G. and Wood, S. N. (2012) Coverage properties of confidence intervals for generalized additive model components. *Scand. J. Statist.*, **39**, 53–74.

Martín-Fernández, J. A., Palarea-Albaladejo, J. and Olea, R. A. (2011) Dealing with zeros. In *Compositional Data Analysis: Theory and Applications* (eds V. Pawlowsky-Glahn and A. Buccianti), pp. 43–58. Chichester: Wiley.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2011) The principle of working on coordinates. In *Compositional Data Analysis: Theory and Applications* (eds V. Pawlowsky-Glahn and A. Buccianti). Chichester: Wiley.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*, 2nd edn. New York: Chapman and Hall.

Paefgen, J., Staake, T. and Fleisch, E. (2014) Multivariate exposure modeling of accident risk: insights from pay-as-you-drive insurance data. *Transprtn Res.* A, **61**, 27–40.

Parry, I. W. H. (2005) Is pay-as-you-drive insurance a better way to reduce gasoline than gasoline taxes? *Am. Econ. Rev.*, **95**, 288–293.

Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosana-Delgado, R. (2015) *Modeling and Analysis of Compositional Data*. Chichester: Wiley.

R Core Team (2017) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reiss, P. T. and Ogden, R. T. (2009) Smoothing parameter selection for a class of semiparametric linear models. *J. R. Statist. Soc.* B, **71**, 505–523.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.

Toledo, T., Musicant, O. and Lotan, T. (2008) In-vehicle data recorders for monitoring and feedback on drivers' behavior. *Transprtn Res.* C, **16**, 320–331.

Tselentis, D. I., Yannis, G. and Vlahogianni, E. I. (2016) Innovative insurance schemes: pay as/how you drive. *Transprtn Res. Proc.*, **14**, 362–371.

Wahba, G. (1981) Spline interpolation and smoothing on the sphere. *SIAM J. Scient. Statist. Comput.*, **2**, 5–16.

Wood, S. (2006) *Generalized Additive Models: an Introduction with R*. Boca Raton: Chapman and Hall–CRC.

Wood, S. N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Statist. Soc.* B, **73**, 3–36.

Wood, S. N. (2013) A simple test for random effects in regression models. *Biometrika*, **100**, 1005–1010.

Wood, S. N., Pya, N. and Säfken, B. (2016) Smoothing parameter and model selection for general smooth models. *J. Am. Statist. Ass.*, **111**, 1548–1563.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Unraveling the predictive power of telematics data in car insurance pricing: Supplementary Material'.